
TRAVEL DEMAND MODELING AND NETWORK ASSIGNMENT MODELS

A PEER-REVIEWED PUBLICATION OF THE TRANSPORTATION RESEARCH BOARD, NATIONAL RESEARCH COUNCIL, NATIONAL ACADEMY PRESS

1994

TRANSPORTATION RESEARCH RECORD 1443

The **Transportation Research Board** is a unit of the National Research Council, which serves the National Academy of Sciences and the National Academy of Engineering. The Board's purpose is to stimulate research concerning the nature and performance of transportation systems, to disseminate the information produced by the research, and to encourage the application of appropriate research findings. The Board's program is carried out by more than 330 committees, task forces, and panels composed of more than 3,900 administrators, engineers, social scientists, attorneys, educators, and others concerned with transportation; they serve without compensation. The program is supported by state transportation and highway departments, the modal administrations of the U.S. Department of Transportation, and other organizations and individuals interested in the development of transportation.

The National Academy of Sciences is a private, nonprofit, self-perpetuating society of distinguished scholars engaged in scientific and engineering research, dedicated to the furtherance of science and technology and to their use for the general welfare. Upon the authority of the charter granted to it by the Congress in 1863, the Academy has a mandate that requires it to advise the federal government on scientific and technical matters. Dr. Bruce M. Alberts is president of the National Academy of Sciences.

The National Academy of Engineering was established in 1964, under the charter of the National Academy of Sciences, as a parallel organization of outstanding engineers. It is autonomous in its administration and in the selection of its members, sharing with the National Academy of Sciences the responsibility for advising the federal government. The National Academy of Engineering, also sponsors engineering programs aimed at meeting national needs, encourages education and research, and recognizes the superior achievements of engineers. Dr. Robert M. White is president of the National Academy of Engineering.

The Institute of Medicine was established in 1970 by the National Academy of Sciences to secure the services of eminent members of appropriate professions in the examination of policy matters pertaining to the health of the public. The Institute acts under the responsibility given to the National Academy of Sciences by its congressional charter to be an adviser to the federal government and, upon its own initiative, to identify issues of medical care, research, and education. Dr. Kenneth I. Shine is president of the Institute of Medicine.

The National Research Council was organized by the National Academy of Sciences in 1916 to associate the broad community of science and technology with the Academy's purpose of furthering knowledge and advising the federal government. Functioning in accordance with general policies determined by the Academy, the Council has become the principal operating agency of both the National Academy of Sciences and the National Academy of Engineering in providing services to the government, the public, and the scientific and engineering communities. The Council is administered jointly by both the Academies and the Institute of Medicine. Dr. Bruce M. Alberts and Dr. Robert M. White are chairman and vice chairman, respectively, of the National Research Council.

The following acronyms are used without definitions in Record papers:

AASHO	American Association of State Highway Officials
AASHTO	American Association of State Highway and Transportation Officials (formerly AASHO)
ASCE	American Society of Civil Engineers
ASME	American Society of Mechanical Engineers
ASTM	American Society for Testing and Materials
FAA	Federal Aviation Administration
FHWA	Federal Highway Administration
FRA	Federal Railroad Administration
FTA	Federal Transit Administration
IEEE	Institute of Electrical and Electronics Engineers
ITE	Institute of Transportation Engineers
NCHRP	National Cooperative Highway Research Program
NCTRP	National Cooperative Transit Research and Development Program
NHTSA	National Highway Traffic Safety Administration
SAE	Society of Automotive Engineers
TRB	Transportation Research Board

TRANSPORTATION RESEARCH RECORDS, which are published on an irregular basis throughout the year, consist of collections of papers on specific transportation modes and subject areas. The series primarily contains the more than 600 papers prepared for presentation at Transportation Research Board Annual Meetings; occasionally the proceedings of other TRB conferences or workshops are also published. Each Record is classified according to the subscriber category covered in the papers published in that volume. The views expressed in the papers are those of the authors and do not necessarily reflect the views of the sponsoring committee(s), the Transportation Research Board, the National Research Council, or the sponsors of TRB activities. The Transportation Research Board does not endorse products or manufacturers; trade and manufacturers' names may appear in a Record paper only because they are considered essential to its object.

PEER REVIEW OF PAPERS: All papers (Annual Meeting papers and those presented at other TRB conferences or submitted solely for publication) in the Transportation Research Records have been reviewed and accepted for publication by the Transportation Research Board's peer review process established according to procedures approved by the Governing Board of the National Research Council. Papers are refereed by TRB technical committees as identified in each Record. Reviewers are selected among committee members and other outside experts. The Transportation Research Board requires a minimum of three reviews; a decision is based on reviewer comments and resultant author revision.

TRANSPORTATION RESEARCH BOARD PUBLICATIONS are available by ordering individual publications directly from the TRB Business Office or by annual subscription through organizational or individual affiliation with TRB. Affiliates and library subscribers are eligible for substantial discounts. For further information or to obtain a catalog of TRB publications in print, write to Transportation Research Board, Business Office, National Research Council, 2101 Constitution Avenue, N.W., Washington, D.C. 20418 (telephone 202-334-3214).

1994 TRANSPORTATION RESEARCH BOARD EXECUTIVE COMMITTEE

Chairman: Joseph M. Sussman, JR East Professor and Professor of Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge
Vice Chairman: Lillian C. Liburdi, Director, Port Department, The Port Authority of New York and New Jersey, New York City
Executive Director: Robert E. Skinner, Jr., Transportation Research Board

Brian J. L. Berry, Lloyd Viel Berkner Regental Professor and Chair, Bruton Center for Development Studies, University of Texas at Dallas
Dwight M. Bower, Director, Idaho Transportation Department, Boise
John E. Breen, The Nasser I. Al-Rashid Chair in Civil Engineering, Department of Civil Engineering, The University of Texas at Austin
Kirk Brown, Secretary, Illinois Department of Transportation, Springfield
David Burwell, President, Rails-to-Trails Conservancy, Washington, D.C.
L. G. (Gary) Byrd, Consultant, Alexandria, Virginia
A. Ray Chamberlain, Vice President of Freight Policy, American Trucking Associations, Alexandria, Virginia (Past Chairman, 1993)
Ray W. Clough (Nishkian Professor of Structural Engineering Emeritus, University of California, Berkeley), Structures Consultant, Sunriver, Oregon
Richard K. Davidson, Chairman and CEO, Union Pacific Railroad, Omaha, Nebraska
James C. DeLong, Director of Aviation, Denver International Airport, Colorado
Delon Hampton, Chairman and CEO, Delon Hampton & Associates, Chartered, Washington, D.C.
Don C. Kelly, Secretary and Commissioner of Highways, Transportation Cabinet, Frankfort, Kentucky
Robert Kochanowski, Executive Director, Southwestern Pennsylvania Regional Planning Commission, Pittsburgh
James L. Lammie, President and CEO, Parsons Brinckerhoff, Inc., New York City
William W. Millar, Executive Director, Port Authority of Allegheny County, Pittsburgh, Pennsylvania (Past Chairman, 1992)
Charles P. O'Leary, Jr., Commissioner, New Hampshire Department of Transportation, Concord
Brig. Gen. Jude W. P. Patin (retired), Secretary, Louisiana Department of Transportation and Development, Baton Rouge
Neil Peterson, former Executive Director, Los Angeles County Metropolitan Transportation Commission, Redondo Beach, California
Darel Rensink, Director, Iowa Department of Transportation, Ames
James W. van Loben Sels, Director, California Department of Transportation, Sacramento
C. Michael Walton, Ernest H. Cockrell Centennial Chair in Engineering and Chairman, Department of Civil Engineering, University of Texas at Austin (Past Chairman, 1991)
David N. Wormley, Dean of Engineering, Pennsylvania State University, University Park
Howard Yersulim, Secretary of Transportation, Pennsylvania Department of Transportation, Harrisburg
Robert A. Young III, President, ABE Freight Systems, Inc., Fort Smith, Arkansas
Mike Acott, President, National Asphalt Pavement Association, Lanham, Maryland (ex officio)
Roy A. Allen, Vice President, Research and Test Department, Association of American Railroads, Washington, D.C. (ex officio)
Andrew H. Card, Jr., President and CEO, American Automobile Manufacturers Association, Washington, D.C. (ex officio)
Thomas J. Donohue, President and CEO, American Trucking Associations, Inc., Alexandria, Virginia (ex officio)
Francis B. Francois, Executive Director, American Association of State Highway and Transportation Officials, Washington, D.C. (ex officio)
Jack R. Gilstrap, Executive Vice President, American Public Transit Association, Washington, D.C. (ex officio)
Vice Adm. Albert J. Herberger, Administrator, Maritime Administration, U.S. Department of Transportation (ex officio)
David R. Hinson, Administrator, Federal Aviation Administration, U.S. Department of Transportation (ex officio)
Gordon J. Linton, Administrator, Federal Transit Administration, U.S. Department of Transportation (ex officio)
Ricardo Martinez, Administrator, National Highway Traffic Safety Administration, U.S. Department of Transportation (ex officio)
Jolene M. Molitoris, Administrator, Federal Railroad Administration, U.S. Department of Transportation (ex officio)
Dave Sharma, Administrator, Research and Special Programs Administration, U.S. Department of Transportation (ex officio)
Rodney E. Slater, Administrator, Federal Highway Administration, U.S. Department of Transportation (ex officio)
Lt. Gen. Arthur E. Williams, Chief of Engineers and Commander, U.S. Army Corps of Engineers, Washington, D.C. (ex officio)

The **Transportation Research Board** is a unit of the National Research Council, which serves the National Academy of Sciences and the National Academy of Engineering. The Board's purpose is to stimulate research concerning the nature and performance of transportation systems, to disseminate the information produced by the research, and to encourage the application of appropriate research findings. The Board's program is carried out by more than 330 committees, task forces, and panels composed of more than 3,900 administrators, engineers, social scientists, attorneys, educators, and others concerned with transportation; they serve without compensation. The program is supported by state transportation and highway departments, the modal administrations of the U.S. Department of Transportation, and other organizations and individuals interested in the development of transportation.

The National Academy of Sciences is a private, nonprofit, self-perpetuating society of distinguished scholars engaged in scientific and engineering research, dedicated to the furtherance of science and technology and to their use for the general welfare. Upon the authority of the charter granted to it by the Congress in 1863, the Academy has a mandate that requires it to advise the federal government on scientific and technical matters. Dr. Bruce M. Alberts is president of the National Academy of Sciences.

The National Academy of Engineering was established in 1964, under the charter of the National Academy of Sciences, as a parallel organization of outstanding engineers. It is autonomous in its administration and in the selection of its members, sharing with the National Academy of Sciences the responsibility for advising the federal government. The National Academy of Engineering also sponsors engineering programs aimed at meeting national needs, encourages education and research, and recognizes the superior achievements of engineers. Dr. Robert M. White is president of the National Academy of Engineering.

The Institute of Medicine was established in 1970 by the National Academy of Sciences to secure the services of eminent members of appropriate professions in the examination of policy matters pertaining to the health of the public. The Institute acts under the responsibility given to the National Academy of Sciences by its congressional charter to be an adviser to the federal government and, upon its own initiative, to identify issues of medical care, research, and education. Dr. Kenneth I. Shine is president of the Institute of Medicine.

The National Research Council was organized by the National Academy of Sciences in 1916 to associate the broad community of science and technology with the Academy's purpose of furthering knowledge and advising the federal government. Functioning in accordance with general policies determined by the Academy, the Council has become the principal operating agency of both the National Academy of Sciences and the National Academy of Engineering in providing services to the government, the public, and the scientific and engineering communities. The Council is administered jointly by both the Academies and the Institute of Medicine. Dr. Bruce M. Alberts and Dr. Robert M. White are chairman and vice chairman, respectively, of the National Research Council.

The following acronyms are used without definitions in Record papers:

AASHO	American Association of State Highway Officials
AASHTO	American Association of State Highway and Transportation Officials (formerly AASHO)
ASCE	American Society of Civil Engineers
ASME	American Society of Mechanical Engineers
ASTM	American Society for Testing and Materials
FAA	Federal Aviation Administration
FHWA	Federal Highway Administration
FRA	Federal Railroad Administration
FTA	Federal Transit Administration
IEEE	Institute of Electrical and Electronics Engineers
ITE	Institute of Transportation Engineers
NCHRP	National Cooperative Highway Research Program
NCTRP	National Cooperative Transit Research and Development Program
NHTSA	National Highway Traffic Safety Administration
SAE	Society of Automotive Engineers
TRB	Transportation Research Board

TRANSPORTATION RESEARCH
RECORD

No. 1443

Planning and Administration

Travel Demand Modeling
and Network
Assignment Models

A peer-reviewed publication of the Transportation Research Board

TRANSPORTATION RESEARCH BOARD
NATIONAL RESEARCH COUNCIL

NATIONAL ACADEMY PRESS
WASHINGTON, D.C. 1994

Transportation Research Record 1443

ISSN 0361-1981

ISBN 0-309-05524-5

Price: \$26.00

Subscriber Category

IA planning and administration

Printed in the United States of America

Sponsorship of Transportation Research Record 1443

**GROUP 1---TRANSPORTATION SYSTEMS
PLANNING AND ADMINISTRATION**

Chairman: Thomas F. Humphrey, Massachusetts Institute of Technology

**Transportation Forecasting, Data, and Economics
Section**

Chairman, Mary Lynn Tischer, Virginia Department of Transportation

Committee on Passenger Travel Demand Forecasting

Chairman: Eric Ivan Pas, Duke University

Bernard Alpern, Moshe E. Ben-Akiva, Jeffery M.

Bruggeman, William A. Davidson, Christopher R. Fleet,

David A. Hensher, Alan Joel Horowitz, Joel L. Horowitz,

Ron Jensen-Fisher, Peter M. Jones, Frank S. Koppelman,

David L. Kurth, T. Keith Lawton, David M. Levinsohn,

Fred L. Mannering, Eric J. Miller, Michael R. Morris,

Joseph N. Prashker, Charles L. Purvis, Martin G.

Richards, Earl R. Ruiter, G. Scott Rutherford, Galal M.

Said, Gordon W. Schultz, Peter R. Stopher, Anti Talvitie, A.

Van Der Hoorn

Committee on Transportation Supply Analysis

Chairman: Hani S. Mahmassani, University of Texas at

Austin David E. Boyce, Yupo Chan, Carlos F. Daganzo,

Mark S. Daskin, Michel Gendreau, Theodore S. Glickman,

Ali E. Haghani, Randolph W. Hall, Rudi Hamerslag, Bruce

N. Janson, Haris N. Koutsopoulos, Chryssi Malandraki,

Eric J. Miller, Anna Nagurney, Earl R. Ruiter, K. Nabil A.

Safwat, Mark A. Turnquist

Transportation Systems Planning Section

Chairman: Bruce D. McDowell, U.S. Advisory Committee on International Relations

Committee on Transportation Programming, Planning and Systems Evaluation

Chairman: Lance Newmann, Cambridge Systematics

Secretary: Dale A. Janik, Illinois department of

Transportation

Abdullah Al-Mogbel, Colin H. Alter, Jon A. Bloom, James L.

Covil, Steve L. Eagan, Joel P. Eittinger, William H.

Goldstein, Robert A. Gorman, Peter L. Hathaway, Roger A.

Herzog, Thomas F. Humphrey, Steen Leleur, Cletus R.

Mercier, Alex E. Metcalf, L. Ray Mikelson, Susan P. Mortel,

Norman G. Paulhus, Jr., Kant Rao, Kumares C. Sinha,

Gerald T. Solbeck, Robert E. Stammer, Jr., Darwin G.

Stuart, Antti Talvittie

Transportation Research Board Staff

Robert E. Spicher, Director, Technical Activities

James A. Scott, Transportation Planner

Transportation Research Record 1443

Contents

Foreword 8

Transportation Network Analysis Techniques for Detailed Travel Forecasts
Cathy L. Chang and David L. Kurth 8

Enhancements to Circulator-Distributor Models for Chicago Central Area Based on Recently Collected Survey Data
David L. Kurth, Cathy L. Chang, and Patrick J. Costinett 28

Using 1990 Census Public Use Microdata Sample to Estimate Demographic and Automobile Ownership Models
Charles L. Purvis 43

Practical Approach to Deriving Peak-Hour Estimates from 24-Hour Travel Demand Models
Charles C. Crevo and Uday Virkud 59

Shopping Trip Chains: Current Patterns and Changes Since 1970
Hyungjin Kim, Ashish Sen, Siim Sööt, and Ed Christopher 71

Estimation of Travel Demand Models with Grouped and Missing Income Data
Chandra Bhat 83

Improved Kalman Filtering Approach for Estimating Origin-Destination Matrices for Freeway Corridors
Nanne J. van der Zijpp and Rudi Hamerslag 100

Introducing "Feedback" into Four-Step Travel Forecasting Procedure Versus Equilibrium Solution of Combined Model

David E. Boyce, Yu-Fang Zhang, and Mary R. Lupa 123

Faster Path-Based Algorithm for Traffic Assignment

R. Jayakrishnan, Wei K. Tsai, Joseph N. Prashker, and Subodh Rajadhyaksha 142

Cost Versus Time Equilibrium over a Network

Fabien Leurent 159

Traffic Assignment Under Environmental and Equity Objectives

Laurence R. Rilett and Christine M. Benedek 176

Application of Dynamic Assignment in Washington, D.C., Metropolitan Area

E. de Romph, H. J. M. van Grol, and R. Hamerslag 191

Multiperiod Network Improvement Model

Chien-Hung Wei and Paul Schonfeld 206

FOREWORD

This volume contains papers focusing on forecasting modeling techniques, traffic assignment methods under various conditions, trip chaining, and a network improvement model for programming network improvements.

Among the various forecasting and modeling papers are two that focus on travel forecasting models and the calibration of refined circulator and distributor models for the city of Chicago central area circulator system. Other papers assess the applicability of the 1990 census Public Use Microdata Sample for regional disaggregate automobile ownership choice models, apply a travel demand model that uses annual traffic count to convert the 24-hr travel demand model output to peak-hour estimates to travel, construct a continuous measure of income from grouped and missing income data used for travel demand models, estimate origin-destination (OD) matrices for freeway corridors by using inter-link induction loop data, and compare new solutions produced by various methods of introducing feedback into the four-step forecasting procedure compared with the equilibrium solution of a model that combines the trip distribution, mode split, and assignment steps.

Transportation Network Analysis Techniques for Detailed Travel Forecasts

CATHY L. CHANG AND DAVID L. KURTH¹

The city of Chicago is in the preliminary engineering final environment impact statement phase of planning for a central area circulator system. Because of the wealth of existing bus and transit service and the amount of activity taking place in the central area, detailed modeling of transit options was required. Described here are enhancements to the transportation network coding and analysis made to travel demand forecasting models necessary to properly model the various transit options. These enhancements include determination of travel speeds on the basis of intersection control and signal timings, explicit coding of transit stops, and detailed multipath transit assignments. Finally the reasonability of applying the detailed network processing techniques in typical regional model applications is discussed.

Chicago's central area is one of the most significant and vibrant activity centers in the Midwest. There were approximately 670,000 employees and 56,600 households in the central area in 1985. This level of activity can be sustained only by use of a variety of transportation alternatives. Chicago is served by several commuter rail lines to the far suburbs, rapid-rail lines to the close-in suburbs, and local and express buses within the city proper. In addition taxis and private automobiles are prevalent in the area. The existing public transportation system is focused on the traditional Loop area defined by the elevated rapid-rail tracks. This area is roughly bordered by Wacker Drive on the north and west, Congress Expressway on the south, and Michigan Avenue on the east, as shown in Figure 1. In this compact eight-by-eight-block area most transit riders are able to walk from their alighting station to their final destination.

Expanded development patterns coupled with ever-increasing congestion have resulted in longer travel times within the central area, which now covers a region stretching from North Avenue on the north to Cermak Road on the south and from Halsted Street on the west to Lake Michigan on the east. It is an area approximately 4 mi. long by 2 mi. wide. As the central area grows in shape and size it is no longer reasonable to expect all travelers to walk from a transit stop or parking location to their final destination. Because the central area is expanding, the concept of a central area circulator, or downtown people mover (DPM), has evolved to provide quick and convenient service within the expanded central area. The proposed system would consist of either an improved bus system or light-rail transit (LRT).

PREVIOUS MODELING EFFORTS

The concept of a central area circulator system has been under study for more than 20 years. The

¹ Barton-Aschman Associates, Incorporated, 820 Davis Street, Evanston, Ill. 60201.

process was formalized in 1989 when the city decided to pursue an Alternatives Analysis/Draft Environmental Impact Statement (AA/DEIS) study for a new transportation system. A detailed model capable of projecting ridership for the extensive transit system in the central area and the proposed alternative network configurations was developed for the Chicago central area AA/DEIS study (1) on the basis of modeling for DPM systems developed for Los Angeles, Miami, and Detroit (2-4).

CHICAGO CENTRAL AREA CIRCULATOR PE/FEIS STUDY

The planning for the locally preferred alternative, an LRT circulator-distributor system, has entered the preliminary engineering/final environmental impact statement (PE/FEIS) phase. On the basis of experience in applying the travel forecasting models developed for the AA/DEIS and the, need for increasingly detailed travel forecasts, a number of refinements to the circulator-distributor modeling process were made:

- Representation of the transit, taxi, and automobile networks was refined.
- Coefficients for the distributor mode-choice model were estimated on the basis of locally collected data.
- Model formulations were revised

The first point, network representation and path-building refinements, is the focus of this paper. The last two points are discussed by Kurth et al. in another paper in this Record.

NETWORK MODELING IMPROVEMENTS

In the past AA/DEIS models were used to test several alternative modes and alignments for the circulator. Although the models produced the necessary forecasts for the AA/DEIS, several areas for refinement were identified through the model application process. Network-related refinements were identified for (a) estimation of automobile, taxi, and bus travel times; (b) coding of bus stop locations; and (c) multipath transit assignment improvements.

For the estimation of AA/DEIS speeds the study area was divided into six large districts, with a representative automobile (or taxi) speed in each district. The average speed was applied across an entire district. This simplified method did not explicitly account for signal delay or vehicle acceleration and deceleration delays. The signal delay contributed a substantial amount to the actual travel time. Alternatively, if a signal was not present, the travel time may be less than that obtained with the average speeds. As the trip distances get longer the actual travel speeds more closely resemble the average travel speed. Since taxi speeds were the same as automobile speeds and bus speeds were a function of automobile speeds, this problem affected all three modes.

Since walk trips were explicitly modeled, the under- or over-estimation of automobile, taxi, and bus travel times for short distances adversely affected the mode-choice models. As a result these models had to adjust for the bias in automobile, taxi, and bus travel times for short distances

through constants on those modes to match observed mode shares. However as trip distances increased, the automobile, taxi, and bus travel times were more accurately modeled in comparison with the walk travel times. This resulted in possible bias in estimations of walk trips for longer interchanges.

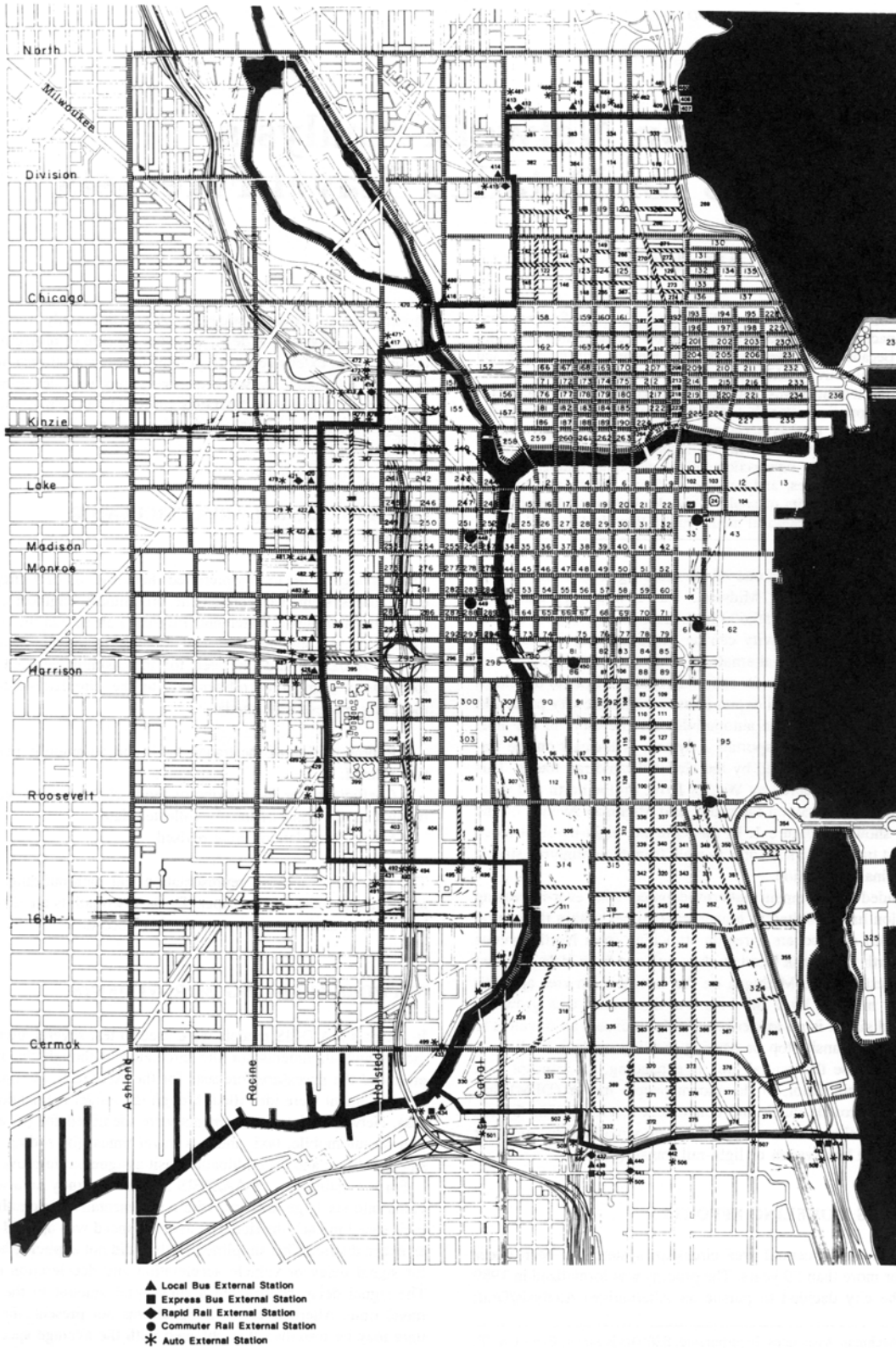


FIGURE 1 Central business district distributor study zone structure.

The second area for refinement was in the coding of the bus stops. In the AA/DEIS study the local buses were allowed to stop at any node (at both cross-street and midblock nodes) along the bus line, although the express buses and cross-Loop buses were modeled to stop only at explicit bus stops on selected "bus-only" lanes. Thus the majority of buses provided ubiquitous bus service, with stops at every node. This was in contrast to the LRT lines, which were coded with explicit stop locations. Thus because of coding conventions buses were modeled to provide more accessible service than the LRT.

The third area identified for refinement was in the modeling of multipath transit assignments. The modeling software used to implement the models, EMME/2, provides a robust algorithm for multipath transit path building as part of its normal procedures. However in the transit-rich environment of the Chicago central area, even the normal EMME/2 transit path-building procedures used in the AA/DEIS tended to underestimate the possible paths for many interchanges. This resulted in some large shifts in transit use on specific lines because of relatively small changes in transit travel times. Thus the most detailed transit path-building algorithm in EMME/2, which is normally reserved for analyzing individual transit interchanges, was used in the PE/FEIS.

The PE/FEIS study provided the opportunity to implement the refinements identified above, along with some commensurate improvements to provide additional detail and sensitivity to the models. The specific network-related refinements included the following:

- Automobile and taxi speeds were estimated by using traffic engineering information, including speed limits, intersection control, signal timing, and link length;
- The bus network was coded in detail, including explicit coding of bus stop location and type (such as near side, far side, and midblock); and
- Bus travel times were built up from link length, bus acceleration and deceleration, number and location of stops, speed limits, and bus dwell times.

The purpose of this paper is to discuss the network modeling improvements that were made in the PE/FEIS procedures, in particular the need for increased detail in coding and the use of transit multipaths.

DESCRIPTION OF NETWORK

In a detailed central business study such as the Chicago circulator project it is imperative to code the network in detail. Detailed decisions regarding track alignments and station locations are made in the PE/FEIS. Changes might be as detailed as moving an alignment or station as little as one block, or 440 ft. If the modeling procedures are not detailed, it is not possible to analyze such changes reasonably.

An integrated highway and transit network was coded for the Chicago central area circulator study area. The highway network was necessary for determining automobile and taxi travel

times, whereas the transit network was needed for determining transit mode-specific travel times. A detailed auxiliary (walk) network was also coded to represent walk paths and transit access and egress. Detailed network coding was important because walking was a viable mode and because of the complexity of the transit system in downtown Chicago. The detailed network coding included, for example, the coding of distances to the nearest one-hundredth of a mile and included all sidewalks (streets) and pedestrian-only links in the network as well as stair links to subway or elevated platforms and the coding of access links from transit stops to the street (auxiliary) network with the equivalent distance to represent the proper travel time from the platform to the street level.

A detailed zone structure was also necessary to properly analyze the trade-offs among walking, taking a taxi, and taking another transit vehicle. Although the detailed zone structure was crucial for improving ridership forecasts, increases in the number of zones increased the difficulty in producing socioeconomic projections for those zones. Thus there was also a practical trade-off restricting the level of detail used in the zone structure. The final zone structure contained 406 internal zones with an 8-Mi² area, as illustrated in Figure 1. Transit external stations were established wherever transit lines crossed the boundary of the study area and at the six central area commuter rail stations. There were 51 mode-specific transit external stations. If several local buses and an express bus crossed the boundary of the study area on the same street, two transit external stations were established—one for the local bus lines and one for the express line. This process prevented spurious transfers between modes at external stations. In addition to transit external stations, 50 automobile external stations were established for possible future use (e.g., performing detailed automobile assignments).

One automobile mode and numerous transit and auxiliary transit modes were used to represent the transportation network. The automobile mode was coded to provide access and egress from zone centroid to zone centroid over the street network. It was used to represent internal-internal automobile travel made by central area residents and internal-internal taxi trip options available to all central area travelers. The transit and auxiliary modes were used in conjunction with another. Auxiliary modes provided access and egress from zone centroids to transit lines and provided for transfer opportunities between nonintersecting transit lines.

Detailed Highway Network Coding and Path Building

The highway network was needed to establish automobile (and taxi) travel times and costs throughout the study area. Automobile and taxi in-vehicle travel times were assumed to be identical. In-vehicle travel times could have been computed by using the AA/ DEIS procedures with average speed zones and link lengths. However this procedure had a tendency to incorrectly estimate automobile or taxi travel times for short trips. To solve this problem it was decided to relate link-specific automobile (and taxi) travel times to traffic engineering information. The total link travel time is composed of the

- *Link traversal time* as a function of link length and speed limit.
- *Delay incurred at an intersection* because of the type of intersection control (5,6).

Different delay functions were developed for signalized intersections, all-way-stop intersections, two-way-stop intersections and Yield signs. The intersection delay for signalized intersections was taken from the 1985 *Highway Capacity Manual* and is a function of the cycle length, green time-to-length (g/c) cycle ratio, volume-to-capacity (v/c) ratio, and lane capacity (7). The four-way-stop function was taken from a report by Meneguzzer et al. (5). The delay associated with a two-way stop was based on the all-way-stop equation with some modifications. The delay, for Yield signs was taken from the function for signals assuming that the yield operated like a signal with a g/c ratio of 1.0.

- *Loss time*, because of vehicle acceleration and deceleration.

To calculate these travel times detailed traffic engineering data were coded on the highway network. This included the following data items for each link:

- Link length.
- Speed limits for all roadways.
- g/c ratios for A.M. and midday for signalized intersections on each approach. and
- Link approach control (e.g., signal, Stop sign, Yield, no control).

The following data items were coded for each node:

- Type of intersection control (signalized, all-way stop, two-way stop, or Yield) and
- Cycle length.

The "congested" automobile and taxi travel times were calculated within the network calculator in EMME/2 by using the travel time functions described above. A set of calibration parameters (average v/c ratios by district) was used to match the modeled automobile speeds with the observed average speeds for 1985. The various components were summed to obtain the total link travel time, which was then stored on the network.

A small number of observed automobile speeds were available for the core area. Although these data were not sufficiently extensive to provide a generalization over the entire study area, they provided a basis for testing the reasonability of the estimated speeds by using the procedures described above. The observed speed data showed substantial variation: speeds on the same road-way rose one year and fell the next year. There was also variation between the morning and evening peak-hour speeds. Most of the peak-hour speeds were between 7 and 10 mph. Table 1 shows a comparison of the observed and modeled automobile speeds for the core area.

Automobile and taxi paths were obtained by running an all-or-nothing assignment on the highway network using the speed information from the link speed calculations. All connector links from zone centroids to the network were coded with 3-mph speeds to represent walk access to the street network. The in-vehicle travel times and travel distances over the shortest time paths were summarized into impedance matrices for use in the mode-choice models.

Taxi fares on the shortest paths were calculated in the matrix calculator on the basis of the

identified as near-side, far-side, or midblock stops. Stair links to the elevated and subway portions of the transit network were included, as were access links representing the distance from commuter-rail platforms to the walk network. The detailed auxiliary (walk) network provided for additional walk access, egress, and transfer between transit lines.

For consistency the bus travel times had to be sensitive to the same traffic engineering information used for obtaining the automobile travel times. Thus the bus travel times were built up in much the same way as the automobile travel times, except that the bus characteristics were used. On the basis of information from CTA, the bus acceleration was set to 1.6 mph/sec, and the bus deceleration was set to 4.7 mph/sec. The bus travel times included the following components:

- Link traversal time,
- Delay incurred at intersections,
- Loss time because of vehicle acceleration and deceleration, and
- Bus stop delay (dwell time)

Bus travel times included a component for bus stop delay that was not included in the automobile travel time calculations. If bus stops were independent of intersections, this delay could simply be estimated and added for each bus stop. However the amount of delay incurred at a bus stop was dependent on the location of a bus stop. For instance near-side bus stop dwell time delay and intersection delay overlap, whereas a far-side bus stop dwell time does not overlap intersection delay.

EMME/2 treats link travel time and transit dwell times independently. Dwell time is coded on the transit network by transit segment (i.e., the portion of the line between bus stops). In addition a transit travel time function is coded on each segment. EMME/2 sums the transit travel time for each link with the dwell time for the segment to determine the total link travel time. Four basic transit travel time functions have been coded on the bus network: one for near-side stops, one for far-side stops, one for midblock stops, and one for no stops.

For near-side bus stops the link travel time function included link traversal time, acceleration and deceleration losses, and the maximum of the dwell time or the signal delay. If greater than zero the increment of dwell time delay over signal delay was also added to the bus travel time.

For far-side bus stops the link travel time function included all of the components of the automobile travel time functions: link traversal time, acceleration and deceleration losses, and signal de-lay time. For the acceleration and deceleration loss, only the deceleration component was included in the calculation because the dwell time for the far-side stop already included both acceleration and deceleration delays.

For midblock stops the bus stop delay was simply added to the traversal time, acceleration and deceleration times, and the signal delay.

Only those buses with a stop were coded with one of the above transit travel time functions. All other segments were coded with a no-stop transit travel time function. This function was

was constructed to represent the acceleration, deceleration, and cruise time between each station as a function of the distance. Acceleration and deceleration rates were provided by the CTA, and a maximum speed of 55 mph was assumed.

For the LRT alternative- the Chicago Circulator Design Team provided the route itinerary, station locations, dwell times, and running times between stations. These time- varied by the configuration of the system, vehicle performance, and operating and safety, considerations.

Transit travel time, were saved by component-transit in-vehicle time, transit auxiliary (walk) time, and transit wait time. The average wait time for a transit vehicle was assumed to be one-half the headway,. A boarding time penalty of 1 min. was used to determine the paths but was not included in the transit impedance matrices.

The disaggregate transit trip analysis techniques embodied within EMME/2 were used to build the transit paths (8). This transit assignment technique is slightly different from the normal EMME/2 transit assignment technique. Both transit path-building techniques build multiple transit piths, but the enhanced path builder provided additional path analysis capabilities that were more suited to a transit-rich environment. This will be discussed in the next section.

Because the mode-choice model is a nested logit model with a local and premium transit choice, two sets of transit paths were necessary: local bus submode and premium bus. Local transit paths were allowed to use local services only. Also since walk paths were explicitly, modeled, the resulting paths were analyzed to ensure that a local transit mode was in fact used. Walk-only strategies were eliminated from local transit paths. Premium transit included the LRT system and any shuttles providing specialized service between specific interchanges and with limited stops in between. An example of this type of bus service is the commuter shuttle serving Illinois Center from the C&NW and Union commuter-rail stations. Again the resulting paths were analyzed to ensure that premium service was indeed used. Walk-only and walk, local bus strategies were eliminated from the premium transit paths. Walk travel times were built by using "sidewalk" links coded at 3 mph.

TRANSIT MULTIPATHS

In a transit-rich environment the construction of multiple transit paths allows for accurate modeling of individual travel behavior. Because there can be multiple transit paths between origin-destination pairs, it is inappropriate to utilize single-path or all-or-nothing path builders. An all-or-nothing path builder constructs only one path on one mode between any two zones. However if competing services are available, travelers may opt to use different paths. It is unlikely that all transit users use the shortest (lowest-impedance) path. Rather if more than one transit path exists between two points, rational travelers will pick the transit vehicle that arrives at their origin first.

The EMME/2 software package has an improved path-building routine based on the concept of transit strategies (8-10). A strategy is a set of rules that allows passage from origin to destination. A strategy is a single element of a transit traveler's choice set. The number and

types of strategies are dependent on the information available to the traveler. In EMME/2 it is assumed that the only information available to the traveler waiting at a node is which line is to be served next. The traveler can then make the decision whether to board the vehicle. If two or more alternative services exist between an origin node and a destination, travelers are assumed to split between the alternative paths in proportion to the frequency of service. This occurs as long as the difference in in-vehicle travel time is less than the difference in headways of the routes.

Consider the example shown in Figure 2. Three alternative paths between Zone 1 and Zone 2 are shown in Figure 2:

- LRT Path A,
- Local bus Path B, and
- LRT Path C.

Normal, non-EMME/2 shortest-transit path-building algorithms would select only LRT Path A between Zones 1 and 2, since Path A has the minimum travel time. The normal EMME/2 path-building techniques would build a strategy by using both LRT Path A and local bus Path B. The normal EMME/2 path builder will build multiple transit paths from a common node. This common-node access location is defined as the node at which the attached connector link leads to the minimum expected travel time to reach a particular destination. Thus the connector link leaving the zone must be on the minimum path. After that, if appropriate, EMME/2 will build multiple transit paths. The split between two or more alternative paths would be based on the relative frequency of service on the paths.

Although the normal EMME/2 transit path builder is an improvement over the shortest-transit path-building technique, LRT Path C is also a reasonable path. However LRT Path C cannot be accessed at a common node with Paths A and B. Thus in EMME/2's normal path-building routine LRT Path C would not be chosen.

EMME/2 has an enhanced path-building technique that is capable of selecting multiple access nodes. The enhanced path-building technique requires additional information regarding reasonable transit access and egress nodes for each interchange and information on how to distribute the trips between the alternative access nodes. In the example shown in Figure 2 the enhanced path builder would select all three paths as reasonable paths between Zones 1 and 2. The selection of an access node or nodes is dependent on several user-defined parameters set within the program (8). The probability for an access node to be chosen is computed by using a simple logit model:

$$P_i = \frac{e^{-\Phi u_i}}{\sum e^{-\Phi u_i}} \quad i \in I_a, j \in I_a \quad (1)$$

where

I_a = set of access nodes,

- u_i = impedance of trip from access node i to destination plus access time from origin coordinates to access node i , and
- Φ = dispersion parameter.

The dispersion parameter is specified by the user. A large value for the dispersion parameter will lead to the selection of only one access node (similar to the normal EMME/2 path-building technique), whereas a small value for the dispersion parameter will tend to split trips more equally among the alternative access nodes. The travel times used to determine the split are the in-vehicle travel time plus the straight-line access time from the origin zone to the respective access nodes.

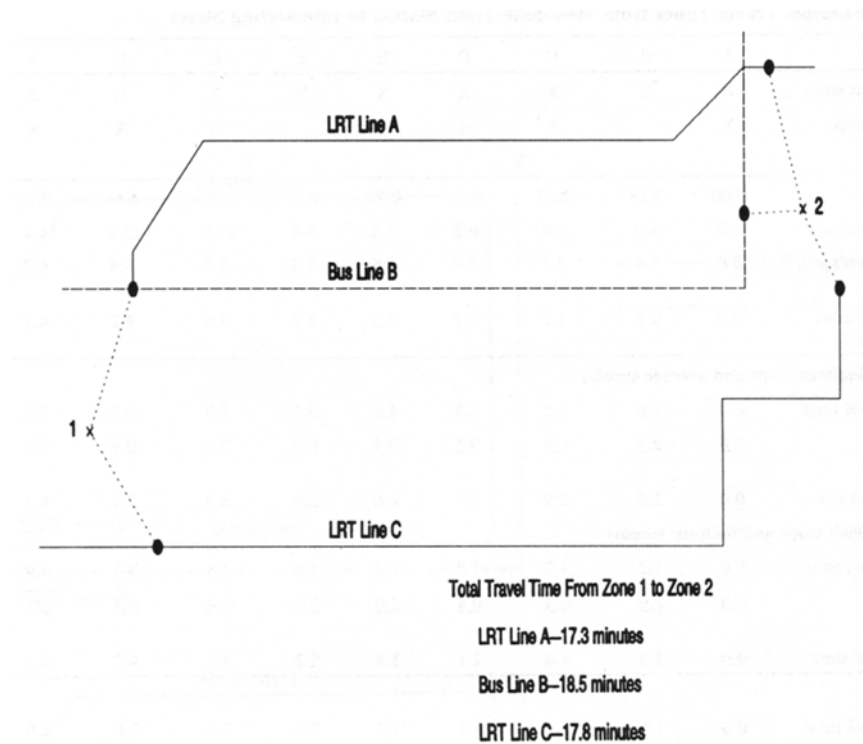


FIGURE 2 Example transit network.

EXAMPLE APPLICATIONS AND COMPARISONS

The network and path-building refinements have been applied in the Chicago central area circulator PE/FEIS study. This section discusses an example application and includes comparisons with the AA/DEIS process. A full comparison of the enhanced network with the procedures used in the AA/DEIS was never performed, since the enhanced procedures were developed to address the observed shortcomings of the AA/DEIS procedures. In addition the choice models used in the PE/FEIS study were updated. This made a direct comparison of the differences in AA/DEIS and PE/FEIS results attributable to network processing changes

impossible.

Network Refinements

The two refinements made in network coding for the PE/FEIS models have been applied in a simple example to show their effects. Travel times were computed by using various network coding schemes for the walk, taxi, bus, and LRT modes for a cross-Loop street running from the west-side train stations. Table 3 shows a comparison of the selected travel times.

The walk travel times were based on an observed walk speed of 3 mph. The average taxi speed was assumed to be 6.5 mph, as determined in the AA/DEIS study. The AA/DEIS bus travel times were based on the 6.5-mph average speed, plus 1.5 min./mi. for dwell time. The LRT was coded in an identical manner for both the AA/DEIS and PE/FEIS studies. The actual bus and LRT stop locations are also shown in Table 3.

In this example the taxi travel times obtained by using the detailed network coding are lower than the travel times obtained by using the 6.5-mph average taxi speed. This occurred because the street modeled had g/c times that favored travel along the street. The average speeds account for travel on "local" streets in the central area as well as major arterials. As can be seen in Table 3, as distances increase, the detailed network travel times approached the taxi travel times calculated using the average taxi speeds.

The bus travel times shown in Table 3 demonstrate the effect of the explicit bus stop coding compared with that of "ubiquitous" stop coding. For the ubiquitous stop coding the walk time is 0.4 min. to all of the intersecting streets. In comparison the walk time for the detailed stop coding varies for each cross street. For example at Street C the walk time is only 0.4 min., since Street C is a bus stop and only the bus access time is represented. However the walk time to Street B is 2.0 min., since it includes the 1.6 min. necessary to walk the 0.08 mi. from the Street C stop back to Street B, in addition to the 0.4-min access time. The same is true for the walk time at the Street D intersection.

This example points out the need to explicitly code bus stops. As can be seen in the example the total travel time to the different cross streets varies substantially depending on the location of the bus stop. Total travel times do not necessarily increase with the distance of the stop from the starting location.

The LRT example shown in Table 3 has characteristics similar to those of the bus coding. Walk times are related to the distance from the closest LRT stop, and in-vehicle travel times are a function of the stop used. The LRT example underscores the importance of explicit coding of bus stops.

TABLE 3 Example Travel Times from West-Side Train Station to Intersecting Street

	A	B	C	D	E	F	G	H	I	J
Ubiquitous bus stops	X	X	X	X	X	X	X	X	X	X
Explicit bus stops	X		X	X			X	X	X	
LRT stations			X				X			
Distance (miles)	0.00	0.15	0.23	0.31	0.39	0.47	0.55	0.64	0.72	0.80
Walk time (minutes)	0.0	3.0	4.6	6.2	7.8	9.4	11.0	12.8	14.4	16.0
Taxi (using average speeds)	0.0	1.4	2.1	2.9	3.6	4.3	5.1	5.9	6.7	7.4
Taxi (using detailed network speeds)	0.0	0.7	1.2	1.7	2.3	3.2	3.6	4.2	4.7	5.4
Bus (with ubiquitous stops and average speeds)										
In-vehicle travel time ^a	0.0	1.6	2.5	3.3	4.2	5.1	5.9	6.9	7.7	8.6
Walk time (stop-specific)	<u>0.0</u>	<u>0.4</u>	<u>0.4</u>	<u>0.4</u>	<u>0.4</u>	<u>0.4</u>	<u>0.4</u>	<u>0.4</u>	<u>0.4</u>	<u>0.4</u>
Total travel time	0.0	2.0	2.9	3.7	4.6	5.4	6.3	7.3	8.1	9.0
Bus (with explicit stops and built-up speeds)										
In-vehicle travel time ^a	0.0	1.2	1.2	1.8	1.8	1.8	3.8	4.3	4.9	4.9
Walk time (stop-specific)	<u>0.0</u>	<u>2.0</u>	<u>0.4</u>	<u>0.4</u>	<u>2.0</u>	<u>3.6</u>	<u>0.4</u>	<u>0.4</u>	<u>0.4</u>	<u>2.0</u>
Total travel time	0.0	3.2	1.6	2.1	3.8	5.3	4.2	4.7	5.3	6.9
LRT										
In-vehicle travel time ^a	0.0	1.3	1.3	1.3	1.3	2.6	2.6	2.6	2.6	2.6
Walk time (station-specific)	<u>0.0</u>	<u>3.0</u>	<u>1.4</u>	<u>1.0</u>	<u>2.6</u>	<u>1.4</u>	<u>1.0</u>	<u>2.4</u>	<u>4.4</u>	<u>5.8</u>
Total travel time	0.0	4.3	2.7	2.3	3.9	3.9	3.5	4.9	6.9	8.3

^aFrom west side train station to nearest transit stop.

The differences between the AA/DEIS and PE/FEIS travel times shown in this example are fairly small-1 to 2 min. Given trips of 3 to 4 mi. the difference could be in the 8- to 10-min range. Also in a detailed analysis in which walking is a viable mode even small changes in the travel times could affect the modal shares. The importance of the travel time differences is even more pronounced, since the coefficients of in-vehicle travel time and walk time are different in the mode choice model.

Path-Building Refinements

A second test was set up within EMME/2 to investigate the effects of different path-building and assignment procedures. The test involved two assignments: the first used the normal EMME/2 transit multipath assignment technique, and the second used the enhanced EMME/2 transit multipath assignment technique. For each of the two assignments 100 trips were assigned to selected interchanges over identical networks. All modes (bus and LRT) were assumed to be available to the travelers. The resulting volumes are shown in Figure 3(a) and (b).

The effect of the enhanced multipath assignment procedure in comparison with that of the normal multipath assignment procedure embodied in EMME/2 is evident if the top diagram in Figure 3 is compared with the diagram below. If the normal and enhanced LRT assignment volumes are compared it can be seen that the assigned volume on the Riverbank LRT route is 349 riders in the enhanced assignment [Figure 3 (*bottom*)] or 151 riders less than the normal multipath assignment [Figure 3 (*top*)]. In the enhanced assignment 56 of the 151 riders are assigned to the Madison Street LRT (the volume increases from 507 in the normal assignment to 563 in the enhanced assignment). The remaining 95 riders are assigned to Madison Street buses.

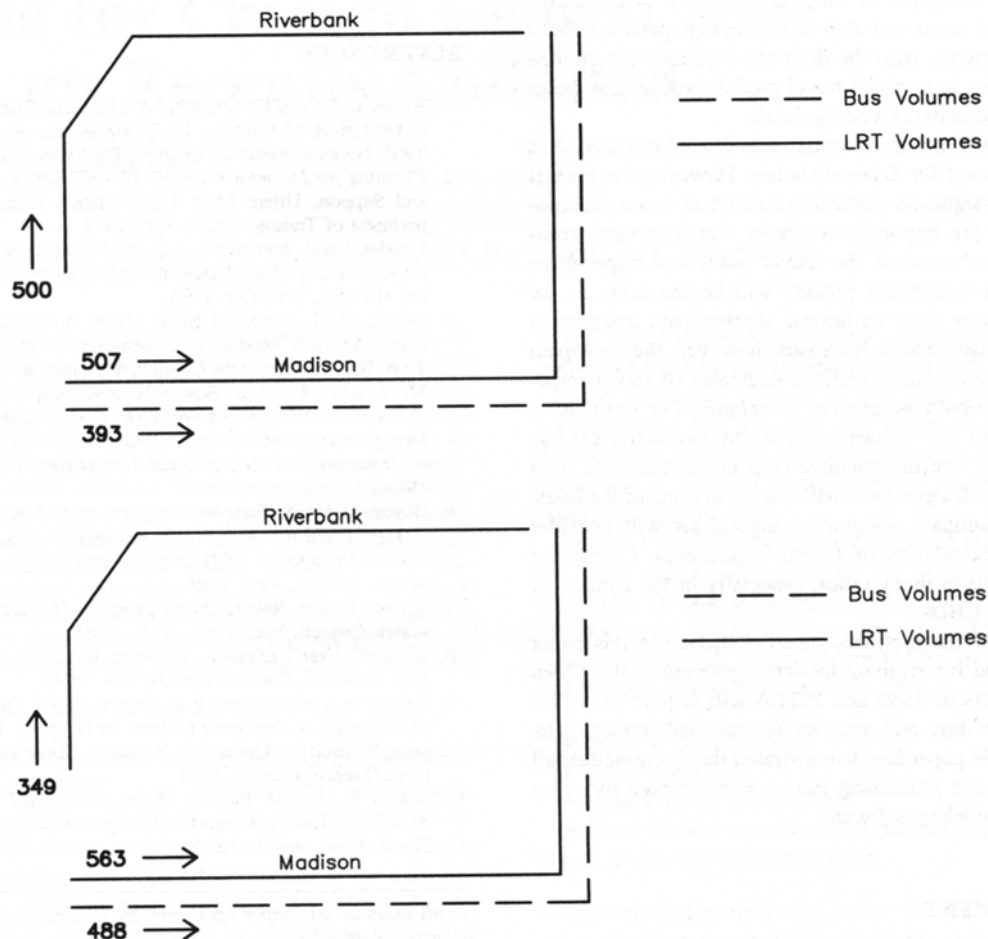


FIGURE 3 Example transit assignment: (*top*) normal and (*bottom*) enhanced travel assignment.

The enhanced multipath assignment process provides additional stability to the assignment process. With the normal multipath assignment process some "flip-flop" of volumes on the Riverbank and Madison Street LRT lines resulted from relatively small changes in travel speeds or route alignments. This occurred whenever the best strategy changed from the use of the station served by the Riverbank LRT line to the station served by the Madison Street LRT line or

vice versa. With the enhanced multipath assignment process changes in travel speeds on one of the lines cause only incremental changes in assigned volumes on the lines.

IMPLICATIONS FOR MODELING PROCEDURES

Several enhancements to the network coding and path-building procedures used for a detailed study of transit alternatives in the central area of Chicago have been presented. The enhancements-calculation of travel times on the basis of intersection control information, explicit coding of bus stops, and detailed transit multipath assignments-were crucial for producing the travel forecasts necessary for station sizing and route design for the PE/FEIS. The examples demonstrated the modeling effort that could be introduced into the process by simplified network coding and processing techniques. Although this effort was necessary for the Chicago central area circulator PE/FEIS, it might be asked whether this level of effort is necessary for regional planning or an AA/DEIS.

It took approximately three times as long as normal to incorporate the detailed traffic engineering information and explicit bus stops. In addition considerable amounts of time and effort were expended in obtaining the data and processing it into a usable format. After this initial investment the time necessary to code alternatives was probably doubled.

One of the major reasons for the increased network detail was the extreme detail necessary for the central area modeling process. Most zones were defined by blocks, the walk mode was explicitly modeled, and taxi, automobile, and bus travel times were directly affected by the location and timing of traffic signals. Most regional modeling processes do not approach the detail of the central area circulator modeling process. Zones typically encompass many blocks, the walk mode is not explicitly modeled (except for bus access, egress, and transfer), and the highway network does not generally include detail for every street in the area being modeled. On the basis of that observation the increased network processing effort described in this paper is probably unwarranted for most regional modeling processes.

Nevertheless some of the enhancements described in this paper should be considered (possibly in a simplified form) for regional modeling processes. The detailed bus stop coding might be appropriate for most modeling processes in central business districts (CBDs). In many regional modeling processes zones in CBDs are typically small, a detailed walk network is generally coded, and the CBD is typically the focus of most transit services and the transit ridership. Thus incorporation of the detailed bus stop coding procedures and possibly the intersection-based travel time calculations in the CBD would be warranted for many regional modeling processes.

In addition the Clean Air Act Amendments of 1990 and Intermodal Surface Transportation Efficiency Act (ISTEA) legislation might also support increased highway network coding and processing detail. The use of average speeds based on, for example, facility type and area type ignores the impact of intersection control on specific streets. This has an impact on the final speeds estimated by the models and subsequently affects air quality calculations. Use of the intersection-based speed estimation procedures described in this paper or simplified versions of those procedures could be an important step in improving speed estimates from traffic

assignments. Also the desire to explicitly incorporate nonmotorized modes in regional travel models would also foster the use of increased network coding detail.

The detailed transit multipath assignment procedures described here are most pertinent for detailed studies. However, the normal transit multipath assignment capabilities (such as those incorporated in EMME/2) are important wherever transit services compete. In many suburban areas the transit paths and impedances from the multipath assignment process will be the same as the paths and impedances from traditional shortest-path assignment techniques. As transit service increases, however, the multipath procedures will produce more realistic estimates of travel impedances than shortest-path assignment techniques. For example in the case in which an interchange is served by two different bus lines that have only one intermediate stop not common to both lines, the shortest-path algorithms will select only one of the lines, whereas normal multipath assignment algorithms will consider both lines in the calculation of transit impedances. Competing transit services occur in many cities, especially in the fringe and urban areas around CBDs.

Although some of the increased detail described in this paper maybe unwarranted for regional modeling processes, the Clean Air Act Amendments of 1990 and ISTEA will require increased detail in network coding and analysis for regional models. The work reported in this paper has demonstrated that increased detail in network coding and processing can be implemented by using readily available modeling software.

ACKNOWLEDGMENTS

The authors would like to thank Robert Kunze of the city of Chicago Circulator Design Office for his support throughout this project. The authors would also like to thank Wayne Miczek and Metra for providing the mode of access survey data to the city of Chicago for use in the calibration of this model. The preparation of this paper was financed in part by the U.S. Department of Transportation, FTA, the state of Illinois Department of Transportation, and the city of Chicago.

REFERENCES

1. Kurth, D. L., and C. L. Chang. A Circulator/Distributor Model for the Central Area of Chicago. In *Transportation Research Record 1328*, TRB, National Research Council, Washington, D.C., 1991.
2. *Planning for Downtown People Movers*. Office of Planning Methods and Support, Urban Mass Transportation Administration, U.S. Department of Transportation, April 1979.
3. Charles River Associates, Inc. *An Independent Forecast of Detroit DPM Ridership*. Joint Center for Urban Mobility Research, Rice Center, Houston, Tex., Oct. 1986.
4. Brooks, K. G., and M.-H. Sung. Miami Downtown People Mover Demand Analysis Model. In *Transportation Research Record 1167*, TRB, National Research Council,

- Washington, D.C., 1988.
5. Meneguzzer, C., D. E. Boyce, N. Rouphail, and A. Sen. *Implementation Evaluation of an Asymmetric, Equilibrium route Choice Model Incorporating Intersection-Related Travel Times*. Illinois Department of Transportation/Chicago Area Transportation Study, Chicago, Sept. 1990.
 6. Boyce, D. E., N. Rouphail, A. Sen, and H. K. Chen. *The Network Design Problem: The Traffic-Responsive Signal Control Scheme*. Illinois Department of Transportation/Chicago Area Transportation Study, Chicago, Aug. 1989.
 7. *Special Report 209: Highway Capacity Manual*. TRB, National Research Council, Washington, D.C., 1985.
 8. *EMME/2 User's Manual-Software Version 6.1*. INRO Consultants, Inc., Montreal, Quebec, Canada, Nov. 1992.
 9. Spiess, H Contributions à la Théorie et aux Outils de Planification des Réseaux de Transport Urbain. Ph. D. thesis. Department of Computer Science and Operations Research, University of Montreal, Montreal, Quebec, Canada, 1984.
 10. Spiess, H., and M. Florian. Optimal Strategies: A New Assignment Model for Transit Networks. *Transportation Research B*, Vol. 23B, No. 2, 1989, pp. 83-102.

Publication of this paper sponsored by Committee on Passenger Travel Demand Forecasting.

The opinions, findings, and conclusions expressed in this paper are not necessarily those of FTA, the Illinois Department of Transportation, or the city of Chicago.

Enhancements to Circulator-Distributor Models for Chicago Central Area Based on Recently Collected Survey Data

DAVID L. KURTH, CATHY L. CHANG, AND PATRICK J. COSTINETT¹

The city of Chicago is evaluating alternative methods of providing for the distribution and circulation of commuters to and workers, visitors, and residents in the vibrant and growing central area of Chicago. In 1990 and 1991 an alternatives analysis/draft environmental impact statement was prepared for a circulator-distributor system for the central area of Chicago. The planning for the locally preferred alternative, a light-rail-transit circulator-distributor system, has now entered the preliminary engineering/final environmental impact statement (PE/FEIS) phase. Refined travel forecasts are being prepared for the PE/FEIS by using refined travel models calibrated with recently collected mode-of-egress survey data. The calibration of the refined circulator-distributor travel models is discussed. In addition the implications for future circulator-distributor and regional modeling efforts that incorporate nonmotorized modes in the choice process are presented.

In 1990 and 1991 an alternatives analysis/draft environmental impact statement (AA/DEIS) was prepared for a circulator-distributor system for the central area of Chicago. Ridership forecasts for the AA/DEIS were prepared by using downtown people mover (DPM) modeling techniques first pioneered for Los Angeles in the early 1970s and later applied in Miami and Detroit (1-3). These models were transferred to the Chicago area and were adjusted to reproduce aggregate travel statistics such as average trip lengths by mode and overall mode shares (4).

The planning for the locally preferred alternative, a light-rail-transit (LRT) circulator-distributor system, has now entered the preliminary engineering/final environmental impact statement (PE/FEIS) phase. On the basis of the experience in applying the travel forecasting models developed for the AA/DEIS and the need for increasingly detailed travel forecasts, a number of refinements to the circulator-distributor modeling process have been made:

- Representation of the transit, taxi, and automobile networks has been refined.
- Coefficients for the distributor mode-choice model have been estimated on the basis of locally collected data.
- Model formulations have been revised.

The last two points are the major focus of this paper. The first point, network representation and path-building refinements, is documented by Chang and Kurth in another paper in this Record.

The travel demand forecasting procedures were applied to a portion of the Chicago region including and surrounding the traditional Loop area (Figure 1). The area modeled encompassed approximately 6.5 mi² and was projected to have more than 83,000 households and 890,000

¹ D.L. Kurth and C.L. Chang, Barton-Aschman Associates, Incorporated, 820 Davis Street, Evanston, Ill. 60201. P.J. Costinett, KJS Associates, Incorporated, 500 108th Avenue, N.E., Suite 2100, Bellevue, Wash. 98004.

employees by 2010. The area is the focus of regional transit services including commuter-rail, rapid-rail, and bus lines.

Figure 1 also shows the detailed zone structure used for the modeling process. Zones within the Loop are generally defined by blocks. Outside the Loop two or more blocks might constitute a single zone. External stations are also defined wherever transit lines cross the study area boundary and for the six major commuter-rail stations included in the study area:

- North Western Station,
- Union Station,
- LaSalle Street Station, and
- Metra Electric commuter-rail stations at Randolph Street, Van Buren Street, and Roosevelt Road.

Two types of internal trips are the primary candidates for travel on a central area circulator-distributor system: internal-internal (circulator) trips and the secondary portion of external-internal and internal-external (distributor) trips. These two types of trips are characterized by marked differences in terms of peaking, activity linkages, regularity, and purpose. Distributor trips are made primarily by central area workers who use regional transit to travel to and from the central area. In the morning these travelers must choose a transit stop at which to leave the transit vehicle that takes them to the central area and the mode of travel (walk, circulator-distributor system, taxi, or a portion of another regional transit route) from the transit stop to the final destination. In the evening the same basic choices are reversed.

In addition to being a major employment and commercial center, the Chicago central area is also a residential area, a cultural center, and a convention center. Thus circulator trip-makers can be divided into several groups on the basis of whether they are residents of the central area, nonresidents of the central area with work as their major purpose for being downtown, or nonresidents of the central area who are downtown for nonwork purposes.

For the Chicago central area the above definitions were used to stratify the travel forecasting model into manageable submodels. Two times of day were explicitly modeled: the morning peak period and midday. Distributor and circulator trips were modeled for both. In the morning peak period the main function of the central area transportation network is the distribution of external-internal trips from regional transit services and commuter rail stations to final destinations. At midday its main function is to provide for central area circulation. The following submodels were developed for forecasting travel within the central area:

- Morning peak period distributor model.
- Morning peak period circulator model for central area residents.
- Midday distributor model.
- Midday circulator model for central area workers.
- Midday circulator model for nonworkers in the central area.
- Midday circulator model for central area residents.

Mode-Choice models were developed for the distributor models for both times of day for all trips entering the central area through one of the six central area commuter-rail stations.

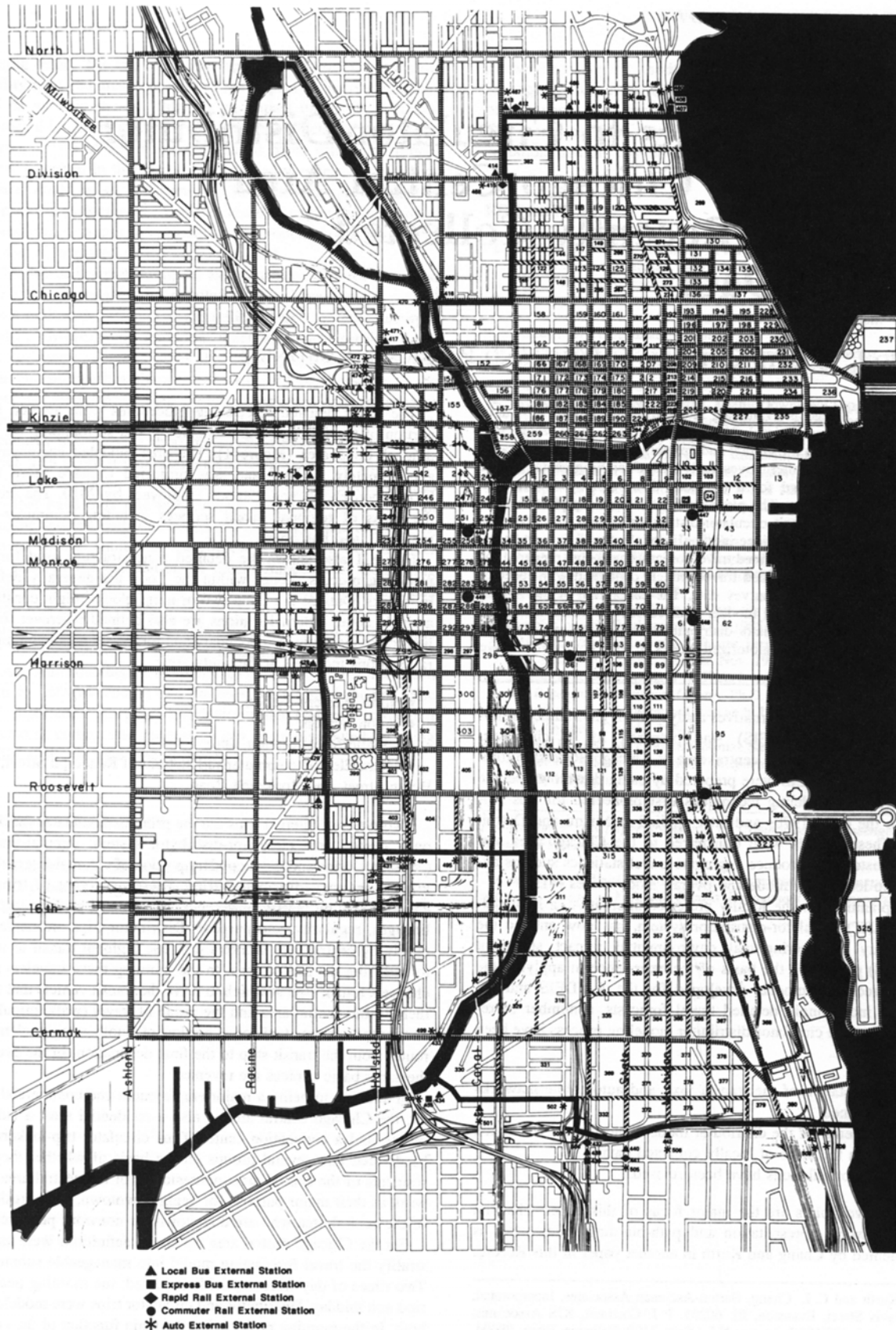


FIGURE 1 Central area circulator zone structure.

The submodes considered at the stations are

- Walk,
- Transit (local bus, express bus, rapid rail, distributor), and
- Taxi.

The submodes provide means to travel from the rail stations to the final destinations in the central area. The trips from commuter-rail stations to final destinations are assigned by submode to their respective networks.

The original DPM models (e.g., for Los Angeles) used a multinomial logit formulation to model mode choice. The modeled distributor systems were "exotic" transit systems such as automated guideway people movers and were considered unique, independent transit modes. The choice alternatives for this model formulation are shown graphically in Figure 2(a).

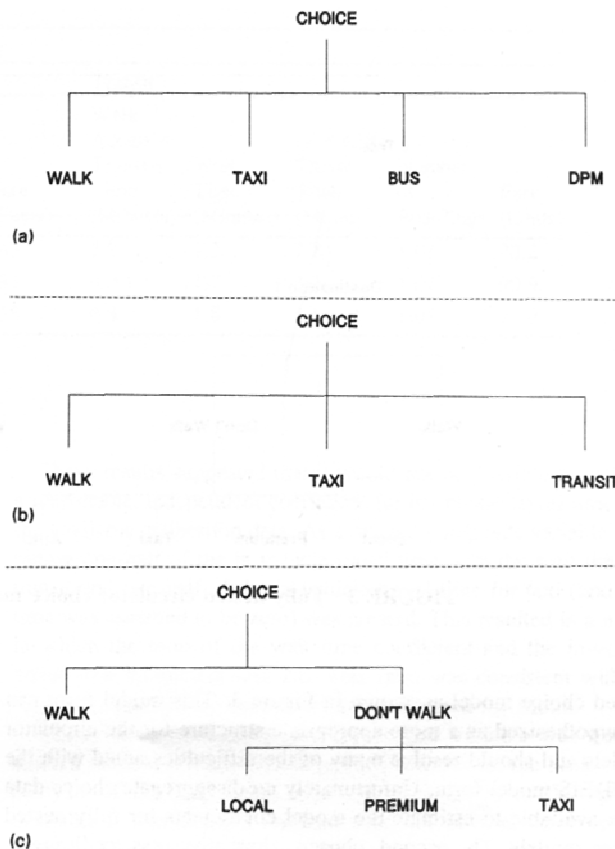


FIGURE 2 Mode-choice model structure: (a) Original DPM multinomial logit (b) AA/DEIS distributor multinomial logit, and (c) PE/FEIS distributor nested-logit.

For the AA/DEIS the choice model was modified to the form shown in Figure 2(b). The distributor alternatives considered for Chicago (transportation system management bus and LRT) were considered to be within the range of transit alternatives already available for distribution

purposes. The distributor was modeled as an alternative path of a generic transit mode rather than as an independent mode.

For the PE/FEIS a nested-logit formulation was used to account for the fact that the proposed alternatives are not truly independent [as in Figure 2(a)], and the use of an LRT distributor system is not the same as riding local buses to final destinations [Figure 2(b)]. The PE/FEIS mode-choice model formulation is shown in Figure 2(c); "local" represents local bus service, and "premium" represents express bus service and LRT.

External-internal trips entering the central area on rapid-rail and bus lines must also be distributed to their final destinations. However unlike trips entering the central area on commuter rail lines, travelers entering the central area are not forced to change their mode at one easily identifiable transit transfer station within the central area. Rather they can ride to the stop nearest their final destination and then walk. Since the transit network in the Chicago central area is so extensive, the distribution of transit riders (i.e., rapid-rail and bus passengers) to their final destinations is accomplished solely through trip assignment techniques. The transit assignment process determines the optimal time paths from "external" transit stations to final destinations and assigns the trips to those paths. The optimal time paths account for in-vehicle travel times, wait times (for transfers), and walk times for transfers and to the final destination.

The estimation of travel in the central area in the circulator mode requires the application of all phases of the travel modeling process: trip generation, trip distribution, mode choice, and trip assignment. Trip generation is based on models developed by Chicago Area Transportation Study (CATS) that generate total person trips, including walk trips, and on the results of a downtown building survey. Trip distribution and mode choice are accomplished through models estimated specifically for the central area. As with the distributor models for trips from commuter-rail stations, circulator trips were assigned to their respective networks by submode. Again the circulator was considered to be part of the premium submode.

A number of observations regarding the simultaneous trip generation, trip distribution, and mode-choice circulator trip modeling methodology used for the AA/DEIS were made. First, the model was difficult to "control." The variables associated mainly with trip distribution interacted with (and sometimes overwhelmed) the mode-choice variables and vice versa. In addition no behavioral explanation could be attributed to the main distribution variable-the natural log of the area of the zone. Finally a matrix balancing technique had to be employed to obtain a reasonable and stable trip distribution.

Two alternatives to the AA/DEIS circulator choice model form were considered for the PE/FEIS model. The first was a fully nested choice model as shown in Figure 3. This model form can be shown in Figure 2(c), with the exception that the premium transit be hypothesized as a more appropriate structure for the circulator submode is replaced by two submodes: express bus and LRT. This model should resolve many of the difficulties noted with the way that was done to allow for the use of a separate mode bias coefficient in the AA/DEIS model form. Unfortunately no disaggregate choice data for LRT for the circulator markets for central area workers and were available to estimate the model coefficients for fully nested central area nonworkers. This procedure is consistent with the choice models. The second, chosen, alternative was to disaggregate the circulator choice models into their component parts and use more traditional sequential modeling process.

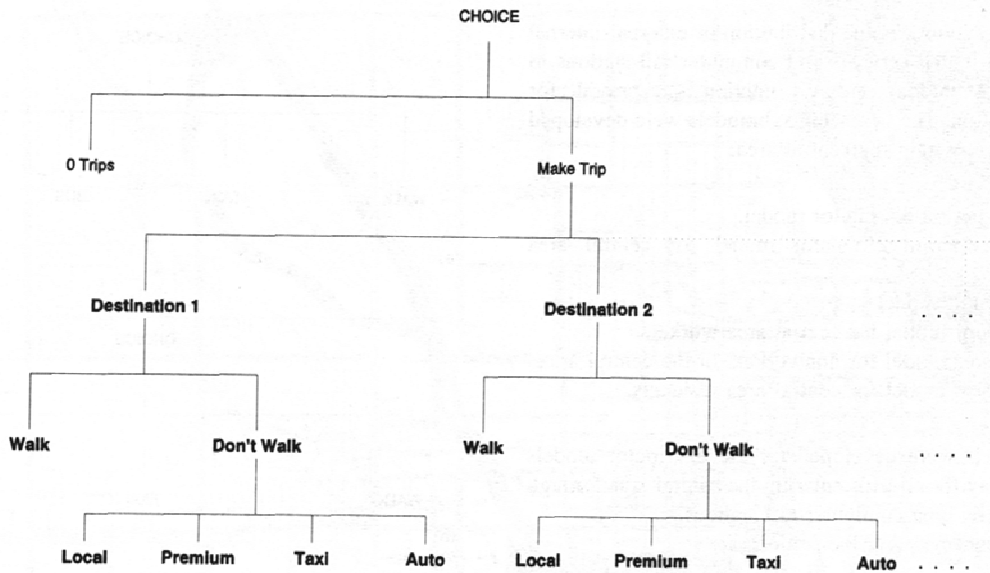


FIGURE 3 Fully nested circulator choice model.

The separation of the simultaneous distribution-mode-choice vehicle over a bus models into their component parts was a drastic change in the modeling methodology. To maintain some impact of the entire transportation system on the trip distribution, the log sum of the mode-choice model was used to define the impedance, or separation, between zones. A traditional gravity model formulation was then used to distribute the trips. Since the original AA/DEIS distribution-mode-choice model included a matrix balancing step to ensure trip attraction balancing in all zones, the conversion to a gravity-type distribution model with composite impedances defined by the denominator of the mode choice model was reasonable.

The circulator mode-choice model form is shown in Figure 4. The model form is very similar to the distributor model form shown in Figure 2(c), with the exception that the premium transit submode is replaced by two submodes: express bus and LRT. This was done to allow for the use of a separate mode bias coefficient for LRT for the circulator markets for central area workers and central area nonworkers. This procedure is consistent with the procedure used in the AA/DEIS and accounts for the hypothesis that, all other travel characteristics being equal, travelers in the central area worker and nonworker markets will select a light-rail vehicle over a bus.

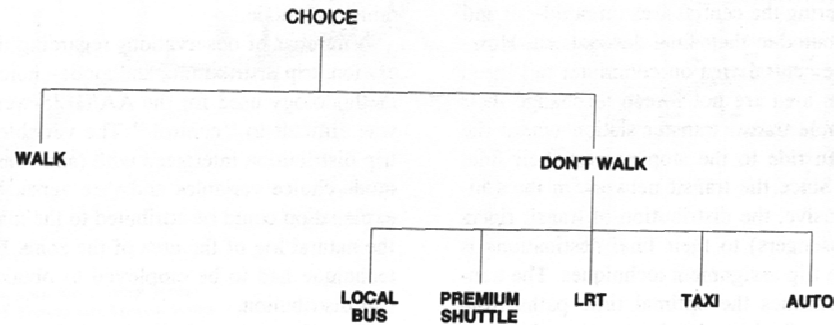


FIGURE 4 Circulator mode-choice model.

CALIBRATION OF PEAK DISTRIBUTOR MODEL

In 1989 Metra performed a mode-of-access survey on commuter-rail lines in the Chicago area (5). The self-administered survey was conducted on the trains and included detailed mode-of-egress and final destination questions. This provided a rich data base of 20,741 individuals observations for the estimation of central area travel models.

Table 1 summarizes the calibration data. The average walk time for walk egress trips was 12.4 min., or about 0.6 mi. This is substantially longer than the 0.44 mi. maximum walk distance used as a rule of thumb in many regional modeling processes. However for the same trips 3.9 min. would be spent, on average, walking to and from taxis, and 5.7 min. would be spent walking to and from transit stops. As would be expected the average walk times (for the walk mode) are substantially higher when taxi or transit was the chosen egress mode.

TABLE 1 Summary of Metra Calibration Data

Mean Values for Alternative Modes											
			Walk	Taxi				Transit			
Chosen Mode	Number of Observations	Percent of Observations	Walk Time (Minutes)	Walk Access Time (Minutes)	In-Vehicle Travel		Walk Access & Transfer Time (Minutes)	Wait Time (Minutes)	In-Vehicle Travel Time (Minutes)	Number of Boardings	Fare (Cents)
				Time (Minutes)	Time (Minutes)	Fare (Cents)					
Walk	9,694	90.3%	12.4	3.9	2.7	142	5.7	1.2	2.8	1.01	90.2
Taxi	109	1.0%	21.7	3.9	4.9	191	6.0	1.7	5.1	1.07	90.9
Transit	938	8.7%	26.2	4.1	5.9	215	4.9	1.8	6.6	1.07	91.0

A logit model estimation program (6) was used to estimate the peak period distributor market mode-choice model. Two preconceived notions guided the calibration. The first was the desire to disaggregate travel time into its component parts-walk time, wait time, and in-vehicle time. The original Los Angeles DPM models used only one travel time variable. This resulted in models that were equally sensitive to changes in walk, wait, or in-vehicle travel times. This situation was modified in the transfer of the models to Chicago for the AA/DEIS through the

addition of a walk distance variable. This variable was necessary to reproduce aggregate mode shares by distance, but since a constant walk speed was used in the modeling process, the variable had the same effect as increasing the walk time coefficient. The second notion was that a nested structure was appropriate for the choice process. The results of the model estimation process led to the final nesting structure used for the peak distributor model [Figure 2(c)].

The final distributor mode-choice model is shown in Table 2 along with the coefficients for models used for the Los Angeles DPM models, the original AA/DEIS study for the Chicago central area circulator, and regional models used in Chicago. It was necessary to create a composite travel time variable for wait time and in-vehicle travel time to obtain a reasonable model coefficient for in-vehicle travel time. All attempts at different model structures that included in-vehicle travel time as an independent variable resulted in positive in-vehicle travel time coefficients. Review of the data summarized in Table 1 provides a reason for the incorrect sign: in-vehicle travel times occur only for the transit and taxi modes, the modes more likely to be used for longer egress trips. Thus the existence of in-vehicle travel time becomes a good variable for explaining why transit or a taxi is used. Both taxi and transit have very similar travel times for the interchanges included in the calibration data set, and taxi has relatively few observations.

TABLE 2 Comparison of Distributor Model Coefficients

Coefficient	Recommended PE/FEIS Model		LA DPM Model	Original AA/DEIS Model	Metra Regional Model	CATS Regional Model ^b
	Coefficient	(t-Score)				
Walk Time	—		-0.09790	-0.2400 ^a	-0.1122	-0.0468
0 - 10 minutes	-0.09152	(-2.8)	—	—	—	—
10(+) - 20 minutes	-0.3461	(-6.0)	—	—	—	—
20(+) - 30 minutes	-0.2385	(-5.6)	—	—	—	—
> 30 minutes	-0.1736	(-4.3)	—	—	—	—
Wait Time	-0.09081	(-1.8)	-0.09790	-0.0900	-0.1122	-0.0173 ^c -0.0290 ^d
In-Vehicle Travel Time	-0.045405	(-1.8)	-0.09790	-0.0900	-0.05611	-0.0159
Travel Cost	-0.01125	(-4.6)	-0.00954	-0.01065	-0.1837	-0.0085
Loop Dummy (on Walk)	0.5600	(3.6)	—	—	0.8843 ^e	—
Nesting Coefficient	0.8943	(6.5)	—	—	0.7064 ^f	—
Constants						
Transit (Local & Premium)	-4.250					
Taxi	-5.380					
Statistics						
Log-Likelihood	-2178.6					
ρ^2 (w.r.t. zero)	0.7843					
ρ^2 (w.r.t. constants)	0.3242					
Value of Time	\$2.42		\$6.16	\$5.07	\$1.83	\$1.12
Year for Dollars	1985		1975	1985	1970	1980 ^g
Value of Time (1985 \$) ^h	\$2.42		\$12.32	\$5.07	\$5.07	\$1.46
Walk / IVTT Ratio	2.0-7.6		1.0	2.67	2.0	2.94
Wait / IVTT Ratio	2.0		1.0	1.0	2.0	1.1-1.8

^aCoefficient on walk distance was converted to time and added to coefficient on walk travel time.

^bFrom CATS regional model for home-based work trips to the Central Business District.

^cFirst wait time.

^dTransfer wait time.

^eFirst nesting coefficient is for lower level sub-mode choice nest and second nesting coefficient is for upper level walk versus drive to transit level nest.

^fConversion to 1985 \$ made using US average CPI-U values.

To test the effect of the lack of difference between the transit and taxi in-vehicle travel times, a special run was performed. The calibration data were modified to reduce the taxi in-vehicle travel time by a factor of 2 for all observations in which a taxi was the chosen mode. This run resulted in the in-vehicle travel time coefficient's being the correct sign and significantly different from zero.

These results suggested that it would not be possible to estimate a reasonable, independent coefficient for in-vehicle travel time with the available calibration data. As a result a composite variable combining one-half of the in-vehicle travel time with the wait time for transit and one-half of the in-vehicle travel time for taxi (taxi wait time was assumed to be zero) was created. This resulted in a model in which the ratio of the wait time coefficient and the in-vehicle travel time coefficient was 2.0. This ratio was consistent with the regional mode-choice model recently calibrated for Metra.

The creation of a composite travel time variable was not the desired method for model estimation. However on the basis of the analysis of the calibration data and an analysis of the options available it was deemed the best solution. Several other options existed. The first would have been to exclude in-vehicle travel time from the model. If this had been done a model with reasonable coefficients for wait time, walk time, and travel cost could have been estimated. It could be argued that the data showed that travelers have little sensitivity to in-vehicle travel time for the portion of their trip from the commuter rail station to their final destination. However, the resulting model would have been valid only for a very limited set of alternatives, since it would not have passed a basic "reasonability" test. Specifically one use of the model will be to test alternative LRT alignments. If in-vehicle travel time is not included in the utility equation, two different alignments would give the same mode choice for a specific interchange as long as walk access and egress distances and headways are the same, even if the in-vehicle travel time of one of the alignments was twice the in-vehicle travel time of the other. Although this example is somewhat illogical, it serves to identify the problem: over what range of travel time differences would the model be valid? A model that excluded in-vehicle travel time as a variable was rejected as illogical.

A second option would have been to transfer a model from a different area. This was the approach used for the AA/DEIS version of the model. That model produced acceptable results for the AA/DEIS study and could possibly have been refined for the PE/FEIS study. It could be argued that this was, in effect, the option chosen. The relationship between the in-vehicle travel time and wait time coefficients was transferred from a regional model estimated by Chicago. Transferring that part of the regional model and rigorously estimating the rest of the model coefficients produced a model more specific and applicable to the Chicago area than transferring a model from another city.

One of the most interesting results of the model calibration was the need to stratify the walk time variable by walk time. The model coefficient for the shortest walk time range, 0 to 10 min., is very similar to the coefficient for wait time. This is consistent with many regional models in which walk and wait times are often grouped into one composite out-of-vehicle travel time variable. The disutility for the second walk time increment, 10 to 20 min., is more than three times as onerous as that for the first walk time increment. Walk times of between 10 and 20 min.

receive the full disutility of walking for 10 min. (i.e., -0.9152) plus the incremental disutility for the portion of the walk greater than 10 min.; walk times of between 20 and 30 min. receive the full disutility for 20 min. (i.e., $-0.09152 \times 10 + -0.3461 \times 10 = -4.3762$) plus the incremental disutility for the portion of the walk greater than 20 min. but less than 30 min., and so on.

TABLE 3 Surveyed and Modeled Mode Shares by Distance

Walk Time Range		Surveyed Shares			Modeled Trips		
Begin	End	Walk	Transit	Taxi	Walk	Transit	Taxi
0	5	96.5%	3.0%	0.5%	99.7%	0.0%	0.3%
5	10	99.1%	0.8%	0.1%	99.6%	0.2%	0.3%
10	15	98.3%	1.3%	0.4%	98.2%	1.4%	0.4%
15	20	91.7%	6.9%	1.3%	92.5%	6.2%	1.2%
20	25	75.0%	22.2%	2.8%	74.9%	22.3%	2.8%
25	30	49.3%	45.6%	5.1%	43.6%	52.6%	3.8%
30	35	31.1%	64.2%	4.7%	21.8%	73.8%	4.4%
35	40	18.9%	81.1%	0.0%	14.1%	81.4%	4.5%
40	45	11.5%	85.2%	3.3%	8.4%	87.8%	3.8%
45	50	0.0%	93.3%	6.7%	6.0%	90.5%	3.4%
50	55	0.0%	100.0%	0.0%	0.0%	100.0%	0.0%
55	60	0.0%	100.0%	0.0%	0.0%	100.0%	0.0%
60	65	0.0%	0.0%	100.0%	0.0%	100.0%	0.0%
65	70	0.0%	100.0%	0.0%	0.0%	100.0%	0.0%

The Loop dummy coefficient is applied to those trips destined to the area bounded by the Chicago River on the north and west, Michigan Avenue on the east, and Congress Parkway on the south. The dummy variable implies that, all other things being equal, travelers are willing to walk longer to destinations inside the Loop than outside the Loop. The willingness of commuters to walk longer distances to Loop destinations is probably an effect of the long history of the traditional Loop area as an employment center served by the existing commuter-rail stations and regular bus service. Historically very little special service (e.g., shuttles) has been provided from the commuter-rail stations to Loop destinations.

The nested model was not statistically significantly better than the root multinomial model with choices between walk, taxi, local bus, and premium transit. The chi-square coefficient comparing the nested model with an equivalent multinomial model (the only difference being the nesting coefficient) was about 0.6. Choosing the nested form did not provide any real improvement in the explanatory power of the model. Nevertheless the nested model was selected since the nesting coefficient was reasonable and the model form fit preconceived notions.

The value of time for the model is about one-half of the value of time for the regional mode-choice model recently calibrated for Metra and for the model used in the AA/DEIS. The value of time was affected by the use of a composite variable to estimate a reasonable in-vehicle travel time coefficient. However the relatively low value of time suggests that commuters are less willing to pay incremental costs to travel from commuter-rail stations to their final destinations.

Table 3 compares the modeled mode shares with the surveyed mode shares by 5-min walk time increments. Figure 5 shows the same information in graphic form. As can be seen in Table 3 and ion Station. Station-specific constants were investigated to improve the results, but they were rejected since their main justification would be to improve the validation results.

receive the full disutility of walking for 10 min. (i.e., -0.9152) plus the incremental disutility for the portion of the walk greater than 10 min.; walk times of between 20 and 30 min. receive the full disutility for 20 min. (i.e., $-0.09152 \times 10 + -0.3461 \times 10 = -4.3762$) plus the incremental disutility for the portion of the walk greater than 20 min. but less than 30 min., and so on.

TABLE 3 Surveyed and Modeled Mode Shares by Distance

Walk Time Range		Surveyed Shares			Modeled Trips		
Begin	End	Walk	Transit	Taxi	Walk	Transit	Taxi
0	5	96.5%	3.0%	0.5%	99.7%	0.0%	0.3%
5	10	99.1%	0.8%	0.1%	99.6%	0.2%	0.3%
10	15	98.3%	1.3%	0.4%	98.2%	1.4%	0.4%
15	20	91.7%	6.9%	1.3%	92.5%	6.2%	1.2%
20	25	75.0%	22.2%	2.8%	74.9%	22.3%	2.8%
25	30	49.3%	45.6%	5.1%	43.6%	52.6%	3.8%
30	35	31.1%	64.2%	4.7%	21.8%	73.8%	4.4%
35	40	18.9%	81.1%	0.0%	14.1%	81.4%	4.5%
40	45	11.5%	85.2%	3.3%	8.4%	87.8%	3.8%
45	50	0.0%	93.3%	6.7%	6.0%	90.5%	3.4%
50	55	0.0%	100.0%	0.0%	0.0%	100.0%	0.0%
55	60	0.0%	100.0%	0.0%	0.0%	100.0%	0.0%
60	65	0.0%	0.0%	100.0%	0.0%	100.0%	0.0%
65	70	0.0%	100.0%	0.0%	0.0%	100.0%	0.0%

The Loop dummy coefficient is applied to those trips destined to the area bounded by the Chicago River on the north and west, Michigan Avenue on the east, and Congress Parkway on the south. The dummy variable implies that, all other things being equal, travelers are willing to walk longer to destinations inside the Loop than outside the Loop. The willingness of commuters to walk longer distances to Loop destinations is probably an effect of the long history of the traditional Loop area as an employment center served by the existing commuter-rail stations and regular bus service. Historically very little special service (e.g., shuttles) has been provided from the commuter-rail stations to Loop destinations.

The nested model was not statistically significantly better than the root multinomial model with choices between walk, taxi, local bus, and premium transit. The chi-square coefficient comparing the nested model with an equivalent multinomial model (the only difference being the nesting coefficient) was about 0.6. Choosing the nested form did not provide any real improvement in the explanatory power of the model. Nevertheless the nested model was selected since the nesting coefficient was reasonable and the model form fit preconceived notions.

The value of time for the model is about one-half of the value of time for the regional mode-choice model recently calibrated for Metra and for the model used in the AA/DEIS. The value of time was affected by the use of a composite variable to estimate a reasonable in-vehicle travel time coefficient. However the relatively low value of time suggests that commuters are less willing to pay incremental costs to travel from commuter-rail stations to their final destinations.

Table 3 compares the modeled mode shares with the surveyed mode shares by 5-min walk time increments. Figure 5 shows the same information in graphic form. As can be seen in Table 3 and ion Station. Station-specific constants were investigated to improve the results, but they were rejected since their main justification would be to improve the validation results.

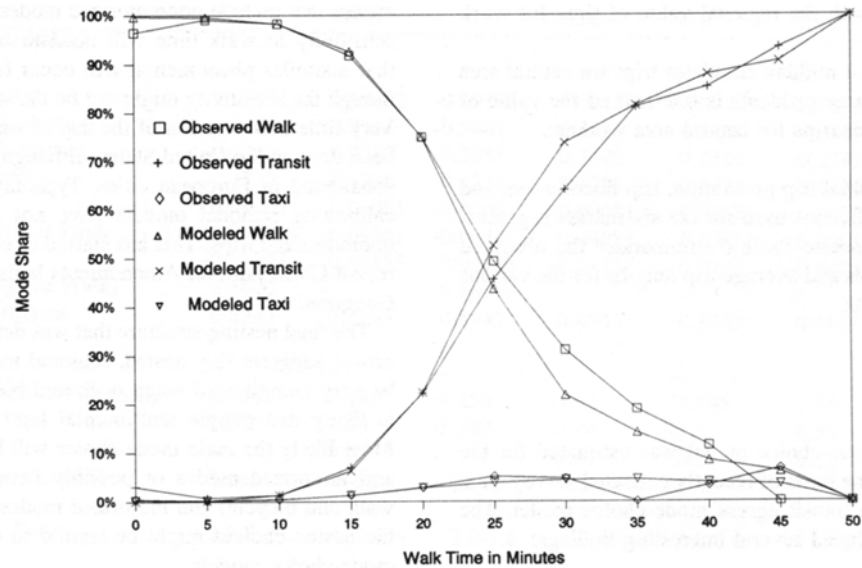


FIGURE 5 Observed and modeled mode shares, peak period distributor trips.

CALIBRATION OF CIRCULATOR MODELS

No disaggregate data existed to rigorously estimate the circulator models. The models were developed on the basis of the relationships determined for the AA/DEIS versions of the models along with the relationships and coefficients determined for the A.M. distributor mode choice models. The assumptions made in the specification of the mode choice model coefficients are summarized below.

- The value of time for A.M. circulation trips for central area residents is comparable with the regional value of time for work trips.
- The value of time for midday circulator trips for central area workers is comparable with the regional value of time for work trips.
- The values of time for midday circulator trips for central area nonworkers and central area residents is one-half of the value of time for midday circulator trips for central area workers.

TABLE 4 Surveyed and Modeled Mode Shares

Station	Surveyed Mode Shares			Modeled Mode Shares		
	Walk	Transit	Taxi	Walk	Transit	Taxi
Van Buren	93.9%	5.1%	1.0%	96.0%	3.4%	0.7%
Randolph	93.1%	5.6%	1.3%	95.9%	3.3%	0.7%
North Western	88.5%	10.6%	0.9%	88.9%	10.1%	1.0%
Union	88.8%	10.1%	1.1%	88.0%	10.9%	1.1%
LaSalle	93.4%	5.6%	1.0%	93.7%	5.3%	1.0%
Total	90.3%	8.7%	1.0%	90.3%	8.7%	1.0%

Table 5 summarizes the final trip generation, trip distribution, and mode-choice model coefficients used for the six market segments used in the modeling process. Table 6 summarizes the observed and estimated mode shares and average trip lengths for the various circulator segment models.

TABLE 5 Trip Generation, Trip Distribution, and Mode-Choice Model Coefficients

Coefficient	AM Peak Distributor Model	AM Peak Circulator Model	Midday Distributor Model	Midday Circulator Model- Workers	Midday Circulator Model- Non- Workers	Midday Circulator Model- Residents
Trip Generation Model^a						
Employment Density (Emp/Ac)—On 0-Trip Util	—	—	—	0.0008552	—	—
Attraction Density (Attr/AC)—On 1-Trip Util	—	—	—	0.00767	—	—
0-Trip Constant	—	—	—	2.75	—	—
Distributor Model^b						
alpha	—	30	—	30	30	30
beta	—	1.10	—	0.30	0.30	0.90
gamma	—	-0.22	—	-0.50	-0.22	-0.60
Distance Coefficients^c						
Walk	—	-7.50	—	-8.37	-6.50	-10.00
Transit	—	-5.50	—	-3.50	-1.00	-4.00
Taxi	—	-4.00	—	-2.00	0.0	-2.00
Auto	—	-4.00	—	-2.00	0.0	-2.00
Mode Choice Model						
Walk Time	—	—	—	—	—	—
0 - 10 minutes	-0.09152	-0.09152	-0.09152	-0.09152	-0.09152	-0.09152
10(+) - 20 minutes	-0.3461	-0.3461	-0.3461	-0.3461	-0.3461	-0.3461
20(+) - 30 minutes	-0.2385	-0.2385	-0.2385	-0.2385	-0.2385	-0.2385
> 30 minutes	-0.1736	-0.1736	-0.1736	-0.1736	-0.1736	-0.1736
Wait Time	-0.09081	-0.09081	-0.09081	-0.09081	-0.09081	-0.09081
In-Vehicle Travel Time	-0.045405	-0.09081	-0.045405	-0.09081	-0.09081	-0.09081
Travel Cost	-0.01125	-0.01125	-0.01125	-0.01125	-0.0225	-0.0225
Loop Dummy (on Walk)	0.5600	—	0.5600	—	—	—
Nesting Coefficient	0.8943	0.8943	0.8943	0.8943	0.8943	0.8943
Constants						
Walk	—	—	—	—	—	—
Transit (Local & Premium)	-4.250	-1.15	-4.250	-1.94	-1.045	-1.20
Taxi	-5.380	-1.40	-5.380	-3.44	-2.5	-2.55
Auto	—	0.0	—	-3.21	0.0	-0.15
Value of Time (1985 \$)	\$2.42	\$4.84	\$2.42	\$4.84	\$2.42	\$2.42
Walk / IVTT Ratio	2.0-7.6	2.0-7.6	2.0-7.6	2.0-7.6	2.0-7.6	2.0-7.6
Wait / IVTT Ratio	2.0	1.0	2.0	1.0	1.0	1.0

^aThe trip generation model utilities are "added" to the composite utilities used for trip distribution. The choice based trip generation model is used only for trips made by CBD workers.

^bThe gamma function has been used to determine friction factors for the gravity model for trip distribution:

$$F = \alpha \times I^{\beta} \times e^{(\gamma \times \gamma)}$$

where:

F is the friction factor for the interchange
I is the composite impedance for the interchange
e is the base of the natural logarithms (2.7183...)
 α , β , and γ are calibrated coefficients

^cThe distance coefficients are applied to the total interchange distance (based on the walk mode shortest travel time paths) and "added" to the composite utilities used for mode choice. This additional utility is used to help control the average trip length by mode.

TABLE 6 Observed and Modeled Mode Shares and Average Trip Lengths, Circulator Model Market Segments

Market Segment		Mode Share				Average Trip Length (Equivalent Walk Minutes) ^a			
		Walk	Transit	Taxi	Auto	Walk	Transit	Taxi	Auto
Peak Circulator—Residents	Observed	51.4%	23.2%	12.6%	12.9%	8.5	21.4	20.1	19.0
	Modeled	51.8%	23.4%	12.2%	12.6%	8.5	25.8	19.3	24.1
Midday Circulator—Workers	Observed	90.1%	6.5%	1.6%	1.7%	4.4	24.7	n/a	n/a
	Modeled	90.1%	6.5%	1.6%	1.7%	4.4	25.6	16.9	23.9
Midday Circulator—Non-Workers	Observed	92.7%	3.5%	0.9%	3.0%	4.4	24.7	n/a	n/a
	Modeled	92.5%	3.5%	0.9%	3.1%	5.3	26.0	10.1	25.5
Midday Circulator—Residents	Observed	92.0%	4.0%	1.0%	3.0%	4.4	24.7	n/a	n/a
	Modeled	91.9%	4.0%	1.0%	3.1%	5.8	23.3	12.2	26.4

^aAll trip lengths are measured using the walk travel times for comparison purposes.

SUMMARY

A detailed distributor mode-choice model was estimated for the Chicago central area on the basis of recently collected survey data. In effect this model is a transit egress mode-choice model. The results of this effort produced several interesting findings:

- A constant value for walk time is not appropriate when the walk time exceeds 10 min. However for walk times of less than 10 min. the disutility of walk time is very similar to the disutility of wait time.
- The implied value of time for the distributor (egress) mode-choice model is about one-half of the value of time for the regional mode-choice model.
- If a nested logit model is used, the proper nesting structure is a choice between "walk" and "don't walk" modes, and between the motorized modes beneath the main "don't walk" mode.

The results of this model calibration effort suggest that future DPM modeling efforts should not be based on the Los Angeles DPM model calibrated in the early 1970s. Although the original model coefficients for travel time and travel cost in Los Angeles are similar to the short walk and wait time and the travel cost coefficients calibrated in the effort described here, the model for Los Angeles did not fully account for the disutility of walking long distances. In addition, the model for Los Angeles probably overestimated the disutility of in-vehicle travel time. Although the likely underestimation of the disutility of long walk time and the overestimation of the disutility of in-vehicle travel time have a tendency to cancel each other in DPM-based models for Los Angeles, they could lead to questionable forecasts of future travel on circulator-distributor or DPM systems.

The results of the present model calibration effort also have implications for future regional modeling efforts that incorporate full mode choice that include nonmotorized modes and for present modeling procedures that include walk access and egress times in the mode choice model. First, when walk time is considered, the disutility of walk time is probably not constant across all time intervals. This study suggests that for times under 10 min. the disutility of walk time is similar to the disutility of wait time. Many existing modeling processes will not suffer, since a general practice has been to limit walk access and egress to 0.33 mi., or about 6.7 min.

However some recent regional modeling efforts have stratified walk access into short walk (less than 0.33 mi.) and long walk (0.33 to 1 mi.). The results of the present study suggest that the coefficient for the long walk access time should be higher than the coefficient for the short walk access time.

When regional modeling efforts begin to incorporate full travel modes that include nonmotorized modes, the effect of varying the sensitivity to walk time will need to be considered. It is likely that a similar phenomenon will occur for bicycle travel time, although the sensitivity might not be the same as that for walk time. Very little investigation of the use of walk and bicycle modes has been done in the United States, although these modes are typically considered in European cities. Typically travel surveys used for calibrating regional models have not collected information on nonmotorized trips. This has started to change, especially with the recent Clean Air Act Amendments legislation passed by the U.S. Congress.

The final nesting structure that was determined for the circulator model suggests that nested, regional mode-choice models might be very complicated when walk and bicycle modes are added. It is likely that simple multinomial logit models will not suffice. More likely the main mode choice will be between walk, bicycle, and motorized modes or possibly between manual modes (i.e., walk and bicycle) and motorized modes. Under motorized modes the nested choices might be similar to those for current regional mode-choice models.

As is typically the case more study and data are required. The current Chicago central area modeling process has been improved by the availability of the Metra mode-of-access and -egress data. However further improvement could be made to the models for the various circulator model segments if comparable data were available for travel made by central area residents, workers, and nonworker visitors. This need will not disappear. It will continue to be necessary as regional planning processes and regional models attempt to consider all travel modes in future modeling efforts.

ACKNOWLEDGMENTS

The authors would like to thank Robert Kunze of the city of Chicago Circulator Design Office for his support throughout this project. The authors would also like to thank Wayne Miczek and Metra for providing the mode of access survey data to the city of Chicago for use in the calibration of this model. The preparation of this paper was financed in part by the U.S. Department of Transportation, FTA, the state of Illinois Department of Transportation, and the city of Chicago.

REFERENCES

1. *Planning for Downtown People Movers.* Office of Planning Methods and Support. Urban Mass Transportation Administration, U.S. Department of Transportation, April 1979.
2. Charles River Associates, Inc. *An Independent Forecast of Detroit DPM Ridership.* Joint Center for Urban Mobility Research, Rice Center. Houston. Tex., Oct. 1986.

3. Brooks, K. G.. and M.-H. Sung. Miami Downtown People Mover Demand Analysis Model. In *Transportation Research Record 1167*, TRB, National Research Council. Washington. D.C., 1988.
4. Kurth. D. L., and C. L. Chang. A Circulator/Distributor Model for the Central Area of Chicago. In *Transportation Research Record 1328*, TRB. National Research Council. Washington, D.C., 1991.
5. The Public Sector Research Group and Market Facts, Inc. *Methodology Report on the Mode of Access Study*. Office of Planning and Analysis, Metra, Chicago, Ill., 1990.
6. Hague Consulting Group. *ALOGIT User's guide Version 3.2*, July 1992.

The opinions, findings, and conclusions expressed in this paper are not necessarily those of FTA, the Illinois Department of Transportation, or the city of Chicago.

Publication of this paper sponsored by Committee on Passenger Travel Demand Forecasting.

Using 1990 Census Public Use Microdata Sample to Estimate Demographic and Automobile Ownership Models

CHARLES L. PURVIS¹

Disaggregate (household-level) automobile ownership choice are typically estimated by using large-scale cross-sectional household travel surveys. Automobile ownership choice models typically stratify households into households owning zero, one, or two or more vehicles. This automobile ownership market segmentation is critical in the application of a regional set of disaggregate travel demand models for aggregate forecasting purposes. An alternative regional data set for estimating disaggregate automobile ownership choice models is the 1990 Census Public Use Microdata Sample (PUMS). PUMS consists of two disaggregate files of individual 1990 census records (household and population characteristics) of either 1 percent of an area's households or 5 percent of an area's households (the 1 percent and the 5 percent samples). Disaggregate workers in household and automobile ownership choice (logit) models were estimated on the basis of PUMS data files for the nine-county San Francisco Bay Area and the one-county San Diego region. These models were also compared with disaggregate models on the basis of the 1990 Metropolitan Transportation Commission household travel survey. The strengths and weaknesses of both approaches-PUMS versus household travel surveys-are discussed. The primary weakness of PUMS is the lack of data on neighborhood characteristics, such as land use density or accessibility measures, at a fine enough geographic level (i.e., regional travel analysis zone) for model estimation purposes. The transferability of the model estimation methodology to other metropolitan regions is discussed.

The purpose of this paper is to explore the development of demographic and automobile ownership forecasting models by using data from the 1990 U.S. decennial census and from household travel surveys. Disaggregate (household-level) automobile ownership choice models were estimated by using data from the 1990 census Public Use Microdata Sample (PUMS) and the 1990 San Francisco Bay Area Household Travel Survey. Comparison of the model estimation results from the two data sets shows that the 1990 census PUMS is an appropriate data set for use in updating metropolitan automobile ownership models. The development of PUMS-based automobile ownership models may be appropriate in metropolitan areas and states where current household travel survey data are not readily available.

TECHNIQUES FOR FORECASTING AUTOMOBILE OWNERSHIP

Travel demand forecasting techniques have typically focused on the four-step planning models related to trip frequency choice, destination choice, mode choice, and route choice. Much less attention is typically paid to the development and evaluation of demographic models that feed data into travel demand models. These demographic models include automobile ownership

¹ Metropolitan Transportation Commission, 101 Eighth Street, Oakland, Calif. 94607-4700.

models, labor force participation rate models, household income models, and age cohort survival models. The focus in this paper is on automobile ownership models, although the point to be made is that the other sets of demographic models are of no less importance. The development of robust and credible labor force participation rate models, household income models, and so on is key to successful urban land use, economic, and transportation stimulation modeling. Demographic and other inputs to travel demand modeling have been covered by Hamburg et al. (1) and Bajpai (2).

Why Is Forecasting Automobile Ownership Important?

Good forecasts of automobile ownership levels are critical in preparing adequate travel demand forecasts. Automobile ownership variables are typically encountered in most travel demand model components, including trip frequency choice, destination choice, and mode choice models.

In terms of trip frequency (trip generation) models, households with no vehicles available take markedly fewer trips than households with one or more vehicles available. Cross-classification or linear regression trip generation (home-based production) models typically include automobile ownership as one of the independent variables used to predict trip frequency choice.

Variables such as the number of automobiles per household, the number of automobiles per worker, and the number of automobiles per licensed drivers have all been used successfully in most if not all work and nonwork mode choice model specifications. Automobile ownership level is less likely to be used in trip destination choice (trip distribution) models, although nested, destination mode choice models invariably include an automobile ownership variable as an independent variable in the mode choice Utility.

Understanding of the numbers of automobiles owned or available to a household and household members is critical in defining the captive market and market choice behavior. Households with no automobiles available will be captive to transit, ride-sharing with nonhousehold members, or nonmotorized means of transportation. Households with multiple workers or drivers per household and only one vehicle per household face a partial captivity-which worker (or driver) gets the family car? Households with one or more cars available per licensed driver and faced with infrequent or inaccessible transit services may essentially be captive or forced to use their automobile because of the lack of alternatives.

Underpredicting future automobile ownership levels will have the effect of underpredicting total motorized person trips, perhaps underpredicting average person trip lengths, overpredicting transit patronage levels, and underpredicting congestion, traffic, and air quality emissions. With these considerations in mind it seems important to get the automobile ownership forecasts right rather than assuming no change in automobile ownership levels with respect to base year automobile ownership levels. The "null model" automobile ownership model (i.e., assuming no change from base year automobile ownership levels) may prove to be an undesirable characteristic of future travel demand model forecasting systems.

Aggregate Versus Disaggregate Automobile Ownership Forecasting Models

Simply stated, aggregate automobile ownership forecasting models are estimated on the basis of areawide time series data on automobile ownership per capita or per household and various independent variables: disaggregate automobile ownership forecasting models are statistically estimated on the basis of household-level data and typically, stratify households into households by the number of automobiles available (e.g., zero, one, two, or three or more automobiles available). Disaggregate automobile ownership models could also be linear regression in mathematical form and would predict the number of automobiles per household, the number of automobiles per capita, the number of automobiles per licensed driver, or the number of automobiles per worker.

Aggregate automobile ownership models can also be estimated by using aggregate zone-level statistics from decennial census data such as the 1980 census Urban Transportation Planning Package (UTPP) or the 1990 census Transportation Planning Package (CTPP). Pearson (3) discusses aggregate automobile ownership models estimated on zone-level data from the 1980 UTPP. Good discussions on aggregate automobile ownership models are included in a publication of the Organization for Economic Cooperation and Development (4). Other relevant discussions on automobile ownership trends and saturation levels are included in reports by Lave (5) and Pisarski (6).

The 1960s state of the practice in disaggregate automobile ownership models is best described by Deutschman (7). These are typically linear regression models predicting automobile ownership rates: the number of automobiles per household or the number of automobiles per capita. Independent variables include average household size, mean or median household income (or log transformations of income), residential density, and single-family versus multifamily dwelling units. Independent variables not analyzed by Deutschman included the numbers of workers in the household and the relative transit accessibility of the residence area with respect to working and shopping opportunities.

Disaggregate automobile ownership rate models (typically linear regression models) can be contrasted with disaggregate automobile ownership level models (typically cross-classification or multinomial logit models). The former predict the number of automobiles per household or the number of automobiles per capita; the latter stratify households by the number of automobiles (or vehicles) owned (or available), say, into categories of zero-vehicle, one-vehicle, and two-or-more vehicle households. Current examples of market-segmented automobile ownership rate models are provided by Prevedouros and Schofer (8). They provide some good exploratory research that may prove to be useful in the formulation of operational, practice-oriented automobile ownership models.

Cross-Classification Automobile Ownership Models

A good example of a cross-classification automobile ownership model is the 1982 version of the Honolulu metropolitan area model (9). The dependent variable is the number of households stratified by three vehicle ownership levels (zero, one, or two or more vehicles per household). Three independent variables are used in the final Honolulu model specification: households by household size (four groups), households by income level (three groups), and households by

geographic area type (three groups). Each of the 36 cells in the cross-classification matrix is assigned three values to split out the shares of households with zero, one, and two or more vehicles. Two other independent variables were examined in the Honolulu analysis: households by number of workers in the household and housing type (single-family versus multifamily units). These two variables were not included in the final model specification, basically to keep the cross-classification model tractable to users. An independent variable not examined in Honolulu included a transit accessibility variable, although one could argue that the area type stratification is perhaps a suitable surrogate for generalized transit accessibility.

Disaggregate Choice Models for Automobile Ownership

Theoretical developments in travel behavior modeling led to the incorporation of nested multinomial logit models to represent automobile ownership choice as a distinct yet integrated element of a "mobility block" of travel demand models [see Lerman (10) and Lerman and Ben-Akiva (11)]. Lerman and Ben-Akiva critiqued the 1970s state of the practice of automobile ownership forecasting as being a "side calculation made with simple models that rely on trend extrapolations or correlations made between 1 and 2 variables and car ownership rather than on a strong causal theory." (A comprehensive review of metropolitan area forecasting models may unfortunately reveal that automobile ownership forecasting is still treated as a "side calculation.")

Two examples of multinomial logit automobile ownership models, in practice, are the Portland, Oregon (12,13), and Bay Area (14,15) travel demand models. Both the Portland and the Bay Area models include a series of mobility block models that first predict the distribution of households by the number of workers in households and then second predict the distribution of households by the number of vehicles in the household. Both the Portland and Bay Area model sets use multinomial logit model specifications to predict the number of workers in households and automobile ownership choice.

The Portland workers-in-household model includes four alternative choices: zero-worker, one-worker, two-worker, or three-or-more-worker households. The utility equations use household size, four income categories, and four categories for age of head of household as independent variables. The Bay Area nonworking household (NWHH) model is a binomial logit model that splits households into households without workers and households with workers. The independent variables included in the Bay Area NWHH model include household size, household income, and special variables to indicate very low income households and low numbers of people per household.

The Portland household automobile ownership model includes four alternative choices: zero-vehicle, one-vehicle, two-vehicle, and three-or-more-vehicle households. Independent variables include the number of households by four household size categories, the number of households by four workers in household categories, the number of households by four income categories, and a generalized transit accessibility variable. This last variable is an "average zonal value of employment accessible within 30 minutes total travel time by transit." Recent revisions to the Portland automobile ownership models, done as part of the 1000 Friends of Oregon Land Use Transportation Air Quality Study, added two variables: the number of retail employees working within 1 mi. of the zone of residence and a "pedestrian environment factor." This last factor is essentially a score assigned to each regional travel analysis zone describing the topography,

sidewalk continuity, local street pattern, and ease of crossing streets within each zone. These urban form variables-employment accessibility and the pedestrian score-help in explaining the lower automobile ownership levels in the central Portland neighborhoods.

The Bay Area has two automobile ownership models-a non-working household automobile ownership (NWHHAO) model and a working household automobile ownership (WHHAO) model. Both Bay Area automobile ownership models split households into the number of households with zero, one, or two or more vehicles available. Independent variables in the current Bay Area NWHHAO model include average household size, average household income, and population density. (The original NWHHAO model included a log sum-based off-peak transit accessibility variable in the model specification). Independent variables in the current Bay Area WHHAO model include average household size, average household income, single-family dwelling unit dummy variable, employment density, and a log sum-based peak transit accessibility variable (essentially a ratio of the exponentiated transit and automobile utilities from the work trip mode choice model). For aggregate model application the Bay Area models are applied to zone-level households market segmented (split) by three household income levels.

The output of the Portland set of worker/automobile ownership choice models is a prediction of the number of households in a travel analysis zone by income groups (four groups), household size (four groups), age of head of household (four groups), numbers of workers in the household (four groups), and number of vehicles available in household (four groups), or essentially up to 1,024 potential market segmentations per zone. The output of the Bay Area set of worker/automobile ownership choice models is a prediction of the number of households in a travel analysis zone by income groups (three groups), number of workers in household (two groups), and number of vehicles available in the household (three groups), or essentially 18 market segmentations per zone. Some of these market segments are likely to be very small in magnitude (e.g., high income, working households with no vehicles available) if not excluded as a potential alternative choice (e.g., three workers in a two-person household).

What are the pros and cons of cross-classification automobile ownership models versus logit choice automobile ownership models? The positive aspects of cross-classification automobile ownership models are their tractability; their ease of specification, estimation, and application; and their ability to satisfactorily handle the highly nonlinear relationships between household income and automobile ownership and between household size and automobile ownership. Readily available data sources for the estimation of cross-classification automobile ownership models include standard census products such as the 1990 CTPP and the 1990 census PUMS. Household travel surveys can also be used for estimating these cross-classification models.

The negative aspects of cross-classification models include a practical (tractable) limitation to two or three independent variables, and aggregation errors related to grouping of what can be considered continuous variables such as household income or residential density. For example a 5 percent increase in mean or median household income in a low-income cohort has no impact on automobile ownership levels in the context of a standard cross-classification automobile ownership model application. In areas with cross-classification automobile ownership models, changes in labor force participation rates, major transit capital investments, or increased development of mixed-use developments and multi-family dwelling units have no impact on automobile ownership forecasts. An alternative to the standard two- or three-dimensional cross-classification model is a more complex-and less tractable-cross-classification automobile

ownership model that could contain four or five independent variables, say, household size, household income level, the number of workers in the household, dwelling unit structure type, and area type or "accessibility class." Large data sets such as the PUMS are ideal for this sort of cross-classification model.

Positive aspects of logit choice automobile ownership models include tractability, ease of estimation and application, and ability to incorporate many of the independent variables that might influence automobile ownership choice. Independent variables that have been included in logit automobile ownership choice models include household size, household income, the number of workers in the household, structure type, employment density and accessibility, transit accessibility to employment, combined transit and highway impedance, population density, and urban design factors. Household travel surveys are the traditional data sources for the estimation of logit choice automobile ownership models. This paper explores the use of the 1990 census PUMS in estimating simple worker/automobile logit choice models.

Negative aspects of logit choice automobile ownership models include the challenges related to model specification, especially with respect to the treatment of the nonlinear relationships between several significant independent variables (e.g., household income and household size) and the dependent variable (the number of households by automobile ownership level). In general logit choice models are less satisfactory in addressing these nonlinear relationships than cross-classification models. In comparison with cross-classification models, logit choice automobile ownership models can be structured to be sensitive to such issues as changes in labor force participation rates, major transit capital investments, and increases in mixed-use land use patterns and multifamily dwelling units.

PUMS AND ITS USE IN TRANSPORTATION PLANNING ANALYSIS

PUMS is a standard Bureau of the Census data product that was first introduced in 1960. The 1990 census version of the microdata sample includes what are called the 1 percent sample and the 5 percent sample as well as a sample of households with elderly householders (16). The PUMS data are basically individual census records for a sample of households and people who answered the census "long form." For example in a region the size of the Bay Area, with 2.246 million households and 6.024 million people, the 5 percent PUMS file for the Bay Area includes disaggregate records on 108,491 households and 292,451 people. This amounts 4.8 percent of the households and 4.9 percent of the total Bay Area population in 1990.

The smallest geographic area for which PUMS data are available is at the Public Use Microdata Area (PUMA). PUMAs may not be less than 100,000 people in total Population in 1990. This large geographic restriction protects the confidentiality of census respondent, by *not* providing precise enough geographic information with which to locate and identify the individual respondent. In 1991 the boundaries of the 1990 census PUMAs were defined by regional census data center staffs as part of the state census data center program. In the nine-county Bay Area, 48 PUMAs each with an average population of 125,000 people were defined.

The PUMS household records include all housing unit data from the 1990 census long form plus recorded variable- such as the number of people in the family and the presence of people age 65

years and over. "Allocation" flag , variables are included to denote if data -values were imputed or allocated by the Census Bureau.

The PUMS person records include person information from the 1990 census long form as well as recorded variables (e.g., recode of place of birth and recode of person's total earnings) and allocation (imputation) flags.

Bay Area transportation planners required 1980 census PUMS data for market segmentation adjustments in the aggregate application of disaggregate choice models (17). Conversion factors were derived from 1980 census PUMS data to convert demographic characteristics of total households into characteristics of households with workers. For example adjustments are needed for four sets of demographic variables included in the Bay Area regional work trip mode choice model TW:

- Income per working household/income per tot.-it household;
- Number of people per working household/number of people per total household;
- Number of workers per working household/number of workers per total household; and
- Number of automobiles per working household/number of automobiles per total household.

Income per working household was 14 percent higher than income per total household, according to the 1980 census PUMS for the Bay Area. Household size in working households was 8 percent higher than household size in total households. The number of workers per working households was 26 percent higher than the number of workers per total household, and the number of automobiles per working household was 11 percent higher than the number of automobiles per total household.

Other PUMS data were used as supplementary data inputs to the Bay Area travel model system to adjust demographic inputs by market segment, namely households by three automobile ownership levels, households by three income levels, and working versus nonworking households.

Summary tabulations for the 1990 census PUMS for the San Francisco Bay Area and the San Diego region are included in Tables 1 through 3 which show the number and characteristics of households by three household income levels and three automobile ownership levels, stratified by total households, working households, and nonworking households, respectively. Information extracted from the 1990 census PUMS is critical for market segmentation adjustments in travel forecasting model systems. These data are also used for the aggregate validation of the workers-in-household model.

Data from the 1990 census PUMS can be charted to show nonlinear relationships between the share of the region's households with no workers in comparison with household income, household size, and age of head of household (Figures 1 through 3, respectively). A graphical exploratory analysis of these demographic relationships assists the model developer in setting up model specifications to properly treat the nonlinear relationships that may appear. Households with a 1989 mean household income of less than \$40,000 have a much higher likelihood of having no workers. One- and two-person households also have a higher likelihood of having no workers. As the age of the head of the household approaches and exceeds 60 years, the likelihood that the household has no workers present increases dramatically.

TABLE 1 Characteristics of Total Households in Bay Area and San Diego 1990 Census PUMS

Total Households by Three Household Income Levels										
	Number and Share of Households				Veh/HH		P/HH		Emp/HH	
	Bay Area	%	San Diego	%	Bay Area	San Diego	Bay Area	San Diego	Bay Area	San Diego
Low Income	783,977	35.0%	378,160	42.7%	1.092	1.236	2.060	2.299	0.737	0.860
Medium Income	791,942	35.3%	316,911	35.8%	1.842	1.995	2.708	2.879	1.482	1.541
High Income	666,635	29.7%	190,503	21.5%	2.439	2.510	3.140	3.171	1.962	1.891
Total	2,242,554	100.0%	885,574	100.0%	1.757	1.782	2.610	2.694	1.364	1.326
Note: "Low Income" = less than \$30,000. "Medium Income" = \$30,000 - \$60,000. "High Income" = greater than \$60,000.										

Note: "Low Income" = less than \$30,000. "Medium Income" = \$30,000 - \$60,000. "High Income" = greater than \$60,000.

Total Households by Three Vehicle Ownership Levels

	Number and Share of Households				Income		P/HH		Emp/HH	
	Bay Area	%	San Diego	%	Bay Area	San Diego	Bay Area	San Diego	Bay Area	San Diego
Zero-Vehicle	235,568	10.5%	71,585	8.1%	\$19,299	\$15,446	1.873	2.117	0.588	0.532
One-Vehicle	729,040	32.5%	296,776	33.5%	\$33,379	\$27,601	1.932	2.039	0.929	0.863
Two-plus Vehicles	1,277,946	57.0%	517,213	58.4%	\$64,603	\$54,594	3.132	3.150	1.755	1.701
Total	2,242,554	100.0%	885,574	100.0%	\$49,693	\$42,384	2.610	2.694	1.364	1.326

TABLE 2 Characteristics of Working Households in Bay Area and San Diego 1990 Census PUMS

Working Households by Three Household Income Levels										
	Number and Share of Households				Veh/HH		P/HH		Emp/HH	
	Bay Area	%	San Diego	%	Bay Area	San Diego	Bay Area	San Diego	Bay Area	San Diego
Low Income	441,614	24.8%	235,687	34.2%	1.292	1.424	2.344	2.620	1.308	1.380
Medium Income	707,365	39.7%	278,915	40.4%	1.885	2.052	2.808	3.003	1.659	1.751
High Income	632,065	35.5%	175,180	25.4%	2.483	2.576	3.210	3.279	2.069	2.057
Total	1,781,044	100.0%	689,782	100.0%	1.950	1.971	2.836	2.942	1.717	1.702
Note: "Low Income" = less than \$30,000. "Medium Income" = \$30,000 - \$60,000. "High Income" = greater than \$60,000.										

Note: "Low Income" = less than \$30,000. "Medium Income" = \$30,000 - \$60,000. "High Income" = greater than \$60,000.

Working Households by Three Vehicle Ownership Levels

	Number and Share of Households				Income		P/HH		Emp/HH	
	Bay Area	%	San Diego	%	Bay Area	San Diego	Bay Area	San Diego	Bay Area	San Diego
Zero-Vehicle	100,411	5.6%	27,052	3.9%	\$28,029	\$21,706	2.277	2.758	1.380	1.408
One-Vehicle	517,763	29.1%	200,106	29.0%	\$37,835	\$30,221	2.058	2.222	1.308	1.280
Two-plus Vehicles	1,162,870	65.3%	462,624	67.1%	\$67,278	\$56,498	3.231	3.265	1.929	1.902
Total	1,781,044	100.0%	689,782	100.0%	\$56,506	\$42,384	2.836	2.942	1.717	1.702

TABLE 3 Characteristics of Nonworking Households in Bay Area and San Diego 1990 Census PUMS

Non-Working Households by Three Household Income Levels										
	Number and Share of Households				Veh/HH		P/HH		Emp/HH	
	Bay Area	%	San Diego	%	Bay Area	San Diego	Bay Area	San Diego	Bay Area	San Diego
Low Income	342,363	74.2%	142,473	72.8%	0.835	0.924	1.693	1.768	0.000	0.000
Medium Income	84,577	18.3%	37,996	19.4%	1.480	1.578	1.865	1.970	0.000	0.000
High Income	34,570	7.5%	15,323	7.8%	1.636	1.755	1.852	1.940	0.000	0.000
Total	461,510	100.0%	195,792	100.0%	1.013	1.116	1.737	1.820	0.000	0.000
Note: "Low Income" = less than \$30,000. "Medium Income" = \$30,000 - \$60,000. "High Income" = greater than \$60,000.										

Note: "Low Income" = less than \$30,000. "Medium Income" = \$30,000 - \$60,000. "High Income" = greater than \$60,000.

Non-Working Households by Three Vehicle Ownership Levels

	Number and Share of Households				Income		P/HH		Emp/HH	
	Bay Area	%	San Diego	%	Bay Area	San Diego	Bay Area	San Diego	Bay Area	San Diego
Zero-Vehicle	135,157	29.3%	44,533	22.7%	\$12,814	\$11,644	1.573	1.729	0.000	0.000
One-Vehicle	211,277	45.8%	96,670	49.4%	\$22,458	\$22,178	1.624	1.660	0.000	0.000
Two-plus Vehicles	115,076	24.9%	54,589	27.9%	\$37,566	\$38,458	2.136	2.178	0.000	0.000
Total	461,510	100.0%	195,792	100.0%	\$23,401	\$24,321	1.737	1.820	0.000	0.000

It is anticipated that transportation planners of the 1990s will use the 1990 census PUMS for various policy and planning analysis, including the following:

- Describing the characteristics of commuters in corridors targeted for congestion pricing programs or major transit or highway capital investments;
- Describing the characteristics of commuting submarkets, including car-poolers, transit passengers, people who work at home, bicycle commuters, disabled people, elderly people and so on;
- Analyzing market segmentation for travel demand forecasting models; and
- Describing the commuting habits by household life cycle stage, occupation of worker, industry, educational attainment, and so on.

The PUMS files are treasure chests of disaggregate household, person, and commuter characteristics that are waiting to be mined by adventurous transportation planners and policy analysts. Needed are case studies to explore conventional and nonconventional ways of using PUMS data to advance transportation planning practice.

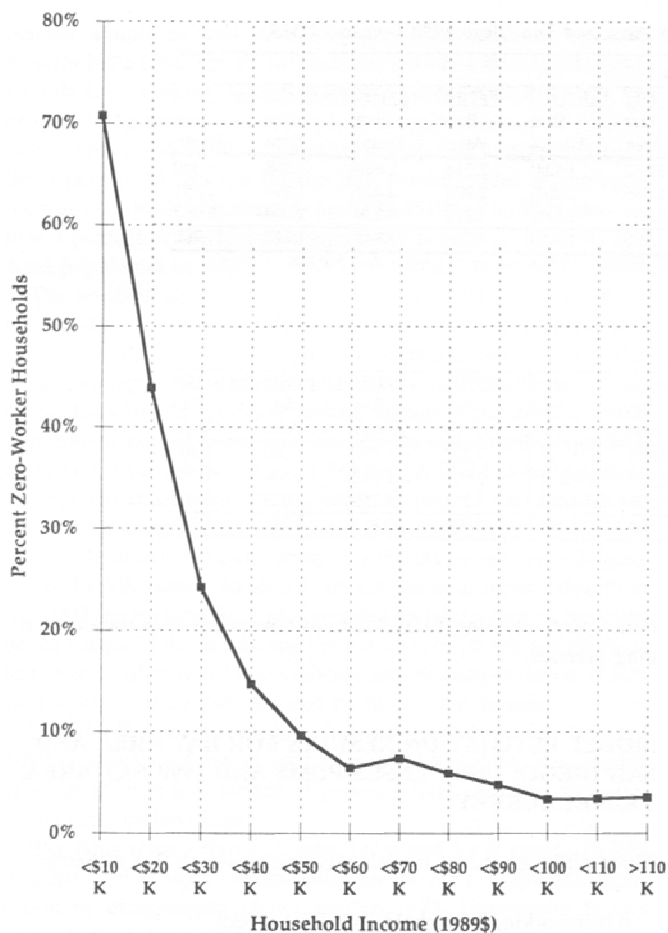


FIGURE 1 Percent zero-worker households by household income ranges, 1989 dollars, 1990 census PUMS, San Francisco Bay Area.

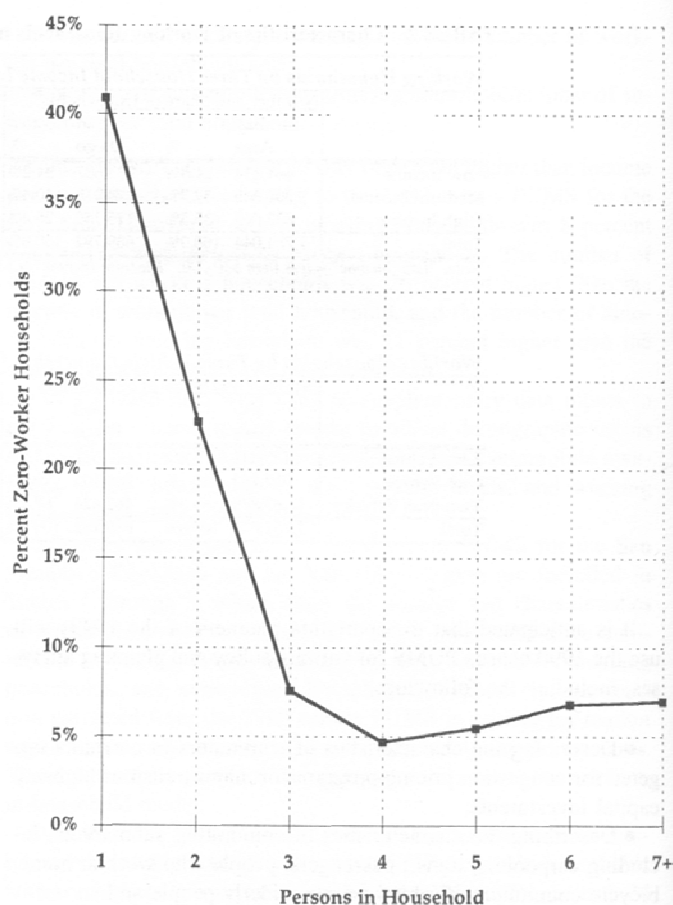


FIGURE 2 Percent zero-worker households by number of people in household, 1990 census PUMS, San Francisco Bay Area.

MODEL ESTIMATION RESULTS FOR BAY AREA AND SAN DIEGO: 1990 CENSUS PUMS AND 1990 BAY AREA TRAVEL SURVEY

Three different choice models were estimated in their research:

- Nonworking household (NWHH) model,
- Nonworking household automobile ownership (NWHHAO) model , and
- Working household automobile ownership (WHHAO) model.

The NWHH model is a binomial logit choice model predicting whether a household has zero or one or more workers. The NWHHAO model is a multinomial logit choice model further splitting nonworking households into households with zero, one, or two or mote vehicles available. The WHHAO model is also a multinomial logit choice model that splits households with workers into households with zero, one, or two or more vehicles available (Figure 4). The models as estimated are similar to previous versions of Bay Area travel demand models, with simplifications and enhancements as noted.

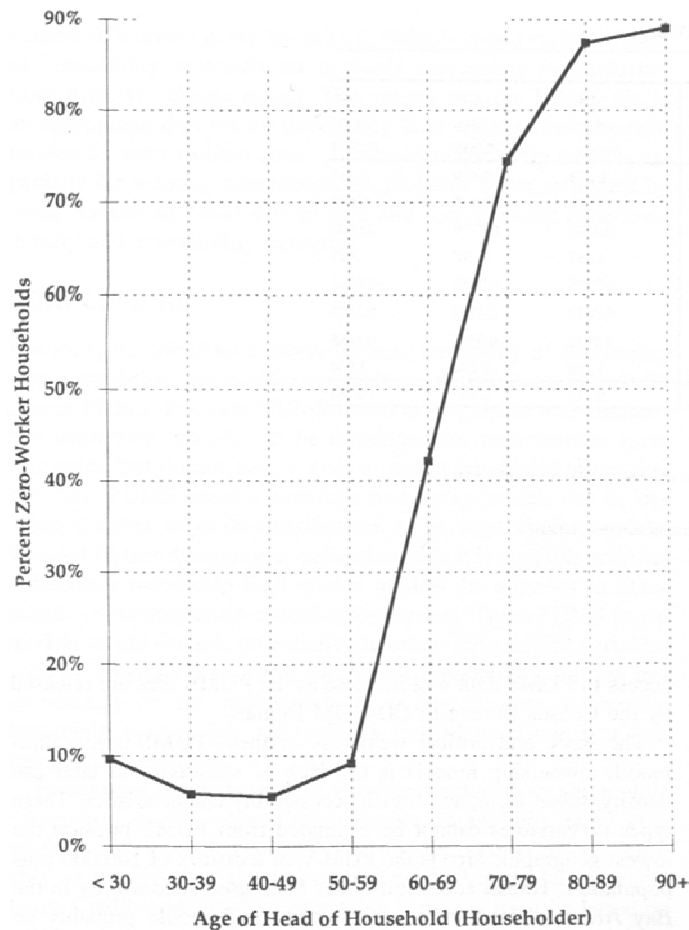


FIGURE 3 Percent zero-worker households by age of head of household, 1990 census PUMS, San Francisco Bay Area.

Six to 10 model specifications were tested for each component model. Only the best model is reported here for the sake of brevity. Three data sets were used in the research project:

- 1990 Bay Area Household Travel Survey,
- 1990 census PUMS 5 percent sample for the Bay Area, and
- 1990 census PUMS 5 percent sample for the San Diego region.

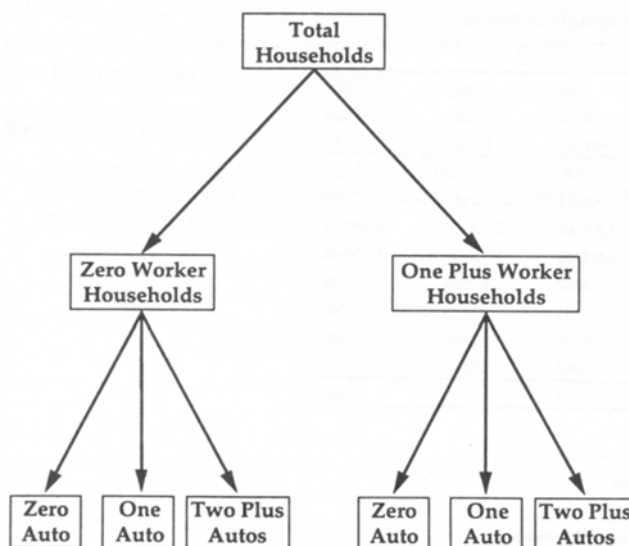


FIGURE 4 Structure of Bay Area automobile ownership models.

The 1990 Bay Area Household Travel Survey was a telephone-based trip diary survey of 10,838 households conducted during the spring and fall of 1990. Households that refused or did not answer the household income question on the 1990 survey (approximately 30 percent of survey respondents) were excluded from the model calibration file. The 5 percent PUMS for the nine-county Bay Area contains 108,491 household records. The 5 percent PUMS for the one-county San Diego region contains 41,987 household records. All PUMS records, including households for which income was imputed, were included in the model specification tests.

Commercially available software was used for preprocessing the rather immense PUMS data sets on a mainframe computer before downloading the calibration files to a microcomputer. Logit models were estimated by using a commercially available logit estimation package.

All of the coefficients were reviewed for reasonableness in terms of coefficient magnitude and sign. All of the t-statistics for all of the coefficients in the summary tables are significant (>2.0) although some of the coefficient signs are counterintuitive. For the present purposes a magnitude of less than a 10-fold difference in model coefficients between data sets is considered a reasonably consistent result (i.e., in the same "ballpark").

The nonworking household model was one of the more difficult models to estimate, although the final rho-bar squared statistics (>0.40) are acceptable for disaggregate choice models (Table 4). The most troublesome variables were the household size variables. The persons per household coefficient for the San Diego PUMS model has the incorrect (negative) coefficient. The low

household size variable (dummy variable of 1 if one-person household and 0 if two-or-more-person household) apparently misbehaves in all models. The income coefficients are quite well behaved and are correct in sign and magnitude. The low-income variable is necessary to correct for the nonlinear relationship between zero-worker household shares and household income. A strong cross-correlation between household size and household income is a main culprit in explaining the incorrect signs for the household size coefficients. The age of the head of household, as well as the average age in the household, was tested in the nonworking household model.

TABLE 4 NWHH Model Specifications

<i>Model #4</i>					
Altern.			Bay	Bay	San
0	1+	Variable	Area	Area	Diego
			Survey	PUMS	PUMS
✓		Constant	-5.546	-5.619	-6.310
	✓	Income	1.32E-05	1.36E-05	9.20E-06
	✓	Persons/HH	0.2035	0.2777	-0.1442
✓		Low Income	1.16E-04	1.05E-04	1.15E-03
✓		Low Pers/HH	-0.108	-0.0539	-0.2478
✓		Age of Head	0.08612	0.08959	0.09252
rho-bar squared			0.427	0.469	0.455

where:

Income = mean household income in 1989 constant dollars.

Persons/HH = persons in household

Low Income = MAX((30000 - Income), 0)

Low Pers/HH = MAX((2 - Persons/HH), 0)

Age of Head = age of head of householder, or householder

The age of the head of household is an extremely powerful variable and basically doubles the rho-bar squared statistics from about 0.24 to 0.43 and above. Inclusion of the age of the head of household variable in the NWHH model raises an important issue: this variable clearly reduces model specification error. On the other hand including the age of the head of household in the model increases model measurement error. How can demographers accurately and confidently forecast the age of the head of household at a zonal level for 20 years into the future? Extensive further disaggregate validation checks are required to determine the value of including age in this demographic model.

With the exception of the household size variables, the estimation results for the NWHH model are quite encouraging when comparing results from the household travel survey with estimation results from the PUMS files for the two California regions.

The nonworking household automobile ownership models show excellent consistency between the Metropolitan Transportation Commission travel survey-based model and two PUMS-based models (Table 5). The household size variable for the one-automobile alternative is rather unstable and probably should be dropped from any final model. The model based on the natural logarithmic transformation of household income seems to work slightly better than that based on mean household income. The single-family dwelling unit dummy variable (1 for single-family, 0

for multifamily) is a strong, intuitive variable that suggests that automobile ownership increases as the share of single-family housing units in a neighborhood increases. The rho-bar squared statistics are quite low (0.16) but are characteristic of multinomial logit choice automobile ownership models.

TABLE 5 NWHHAO Model Specifications

Model #5						
Alternative				Bay Area Survey	Bay Area PUMS	San Diego PUMS
0	1	2+	Variable			
✓			Constant (0)	6.453	5.439	5.861
		✓	Constant (2+)	-9.988	-9.513	-9.673
	✓		Log Income	0.8108	0.5779	0.6900
		✓	Log Income	1.567	1.297	1.445
	✓		Persons/HH	-0.2815	0.0453	-0.0562
		✓	Persons/HH	0.3622	0.4723	0.3703
	✓		Single Family DU	0.7166	0.9588	0.8685
		✓	Single Family DU	1.796	2.290	1.934
rho-bar squared				0.158	0.163	0.161

where:

Log Income = natural logarithm of mean household income in 1989 constant dollars.

Persons/HH = persons in household.

Single Family DU = single family dwelling unit dummy variable, where:

1, if single family dwelling unit.

0, if multiple family dwelling unit.

The working household automobile ownership models are structured similarly to the NWHHAO models (Table 6). The household size variable in the one-automobile utility was dropped because of counterintuitive (negative coefficient) results. Mean household income was used instead of the logarithm of household income. The number of workers per working household variable was added to the two-or-more automobile utility equation to show the impact of multiworker households on increasing the probability of owning two or more vehicles. All coefficient signs are correct in direction. Coefficient magnitudes tend to fluctuate more in this model than in the NWHHAO model, but all coefficients are in the same ballpark.

The model estimation results are basically encouraging and show the utility of using unconventional data sources, that is, the 1990 census PUMS, in statistically estimating selected demand models for a regional travel forecasting system. Results are generally consistent when comparing survey-based models with PUMS-based models and when comparing PUMS-based models between different metropolitan areas. The prospects for the transferability of these models and methodologies to other metropolitan areas are quite good. The San Diego PUMS models are quite similar to the Bay Area PUMS models. Prospects for the ease of access to PUMS data will increase as the PUMS Mes are released by the Census Bureau in CD-ROM format.

TABLE 6 WHHAO Model Specifications

Model #7.5				Bay Area Survey	Bay Area PUMS	San Diego PUMS
Alternative						
0	1	2+	Variable			
✓			Constant (0)	-1.384	-0.736	-1.139
		✓	Constant (2+)	-3.451	-2.951	-2.973
	✓		Income	4.22E-05	2.12E-05	3.03E-05
		✓	Income	6.02E-05	3.95E-05	5.60E-05
		✓	Persons/HH	0.3902	0.2908	0.2232
	✓		Single Family DU	1.2740	0.7576	0.2742
		✓	Single Family DU	2.234	2.070	1.341
		✓	Workers/HH	0.998	0.812	1.015
rho-bar squared				0.221	0.245	0.235

where:

Income = mean household income in 1989 constant dollars.

Persons/HH = persons in household

Single Family DU = single family dwelling unit dummy variable, where:

1, if single family dwelling unit.

0, if multiple family dwelling unit.

Workers/HH = employed residents in household

The basic and critical weakness of these PUMS-based automobile ownership models is the lack of sensitivity to land use density, urban form, and transit accessibility characteristics. These types of variables cannot be estimated from PUMS because the lowest geographic area is the PUMA, or a district of 100,000-plus population. Transit accessibility has been used successfully in the Bay Area and Portland model systems and should probably be incorporated (or at least attempted to be incorporated) into travel forecasting models in large metropolitan areas with significant transit ridership levels and significant shares of zero-automobile households.

The NWHH model is essentially a demographic model that splits households in a travel analysis zone into households by the number of workers in the household. Neighborhood variables such as accessibility or density are probably appropriate for exclusion from a worker choice model. This means that the PUMS file is an appropriate data set for developing final workers-in-household models for metropolitan areas. Automobile ownership models, especially for working households, are probably better estimated by using household travel survey data and incorporating zone-level density and accessibility measures.

CONCLUSIONS

The study described here shows the ease and utility of developing demographic and automobile ownership models by using the 1990 census PUMS data sets. PUMS-based demographic and automobile ownership models can be developed for metropolitan areas and states that do not have access to recent household travel survey data. PUMS-based automobile ownership models can be developed either as cross-classification or as logit choice models. Without further

disaggregate validation tests it is arguable whether automobile ownership logit choice models are superior to automobile ownership cross-classification models. These PUMS-based models would exclude potentially important independent variables such as density, urban form, and transit accessibility, but are improvements on the rudimentary cross-classification automobile ownership models that are typically limited to just one or two independent variables.

Forecasting the independent variables included in automobile ownership models-household income, household size, age, workers in household, accessibility, density, and so on-is arguably as important as the automobile ownership model specifications. Credible automobile ownership forecasts must be based on credible forecasts of the necessary demographic input variables. Research on the development of forecasting models for demographic variables is equally as important as travel behavior research.

Final, revised Bay Area travel demand models will be based on models estimated from the 1990 Bay Area Household Travel Survey rather than the 1990 census PUMS. Final Bay Area travel demand models will build on the insight gleaned from this PUMS-based analysis. A basic conclusion is that the 1990 census PUMS is a "second-best" data set for demographic and automobile ownership model development and is no substitute for a comprehensive household travel behavior survey.

REFERENCES

1. Hamburg, J. R., G. T. Lathrop, and E. J. Kaiser. *NCHRP Report 266: Forecasting Inputs to Transportation Planning*. TRB, National Research Council, Washington, D.C., 1983.
2. Bajpai, J. N. *NCHRP Report 328: Forecasting the Basic Inputs to Transportation Planning at the Zonal Level*. TRB, National Research Council, Washington, D.C., 1990.
3. Pearson, D. F. Disaggregating Zonal Households by Size, Income and Auto Ownership. Presented at Third National Conference on Transportation Planning Methods Applications, Dallas, Tex., April 1991.
4. Organization for Economic Cooperation and Development. *Forecasting Car Ownership and Use*. OECD Road Research Group, Paris, 1982.
5. Lave, C. *Things Won't Get a Lot Worse: The Future of U.S. Traffic Congestion*. UCI-ITS-WP-90-2. Institute of Transportation Studies, University of California, Irvine, 1990.
6. Pisarski, A. E. *Commuting in America: A National Report on Commuting Patterns and Trends*. Eno Foundation for Transportation, Incorporated, Westport, Conn., 1987.
7. Deutschman, H. D. Auto Ownership Revisited: A Review of Methods Used in Estimating and Distributing Auto Ownership. In *Highway Research Record 205*, HRB, National Research Council, Washington, D.C., 1967, pp. 31-49.
8. Prevedouros, P. D., J. L. Schofer. Factors Affecting Automobile Ownership and Use. In *Transportation Research Record 1364*, TRB, National Research Council, Washington, D.C., 1992, pp. 152-160.

9. Schimpeler Corradino Associates. *Oahu Model Update Study: Final Report.- Volume II.* Oahu Metropolitan Planning Organization, Honolulu, Hawaii, 1982.
10. Lerman, S. R. Location, Housing, Automobile Ownership, and Mode to Work: A Joint Choice Model. In *Transportation Research Record* 610, TRB, National Research Council, Washington, D.C., 1976, pp. 6-11.
11. Lerman, S. R., and M. E. Ben-Akiva. Disaggregate Behavioral Model of Automobile Ownership. In *Transportation Research Record* 569, TRB, National Research Council, Washington, D.C., 1976, pp. 34-51.
12. Lawton, T. K. *Travel Forecasting Methodology Report (Regional Models).* Metropolitan Service District, Portland, Oreg., 1989.
13. Cambridge Systematics, Incorporated. *Making the Land Use Transportation Air Quality Connection: Model Modifications: Volume 4.* 1000 Friends of Oregon, Portland, Oreg., November 1992.
14. Ruiter, E. R., and M. E. Ben-Akiva. Disaggregate Travel Demand Models for the San Francisco Bay Area: System Structure, Component Models, and Application Procedures. In *Transportation Research Record* 673, TRB, National Research Council, Washington, D.C., 1978, pp. 121-128.
15. Kollo, H. P. H. *Home-Based Work Trips Models-Final Disaggregate Version: Travel Model Development with 1980181 Data Base.* Working Paper 2. Metropolitan Transportation Commission, Oakland, Calif., 1987.
16. *Public Use Microdata Sample Technical Documentation.* Census of Population and Housing, 1990. U.S. Bureau of the Census, Washington, D.C., 1992.
17. *A Disaggregate Work-Trip Mode Choice Model for Aggregate Forecasting (Model TW): Technical Summary: Travel Model Update with 1980181 Data Base.* Metropolitan Transportation Commission, Oakland, Calif., April, 1988.

Publication of this paper sponsored by Committee on Passenger Travel Demand Forecasting.

Practical Approach to Deriving Peak-Hour Estimates from 24-Hour Travel Demand Models

CHARLES C. CREVO AND UDAY VIRKUD¹

The Clean Air Act Amendments of 1990 have created a need for accurate and reliable estimates of peak-period travel and the speeds at which vehicles operate during that time period. State transportation agencies are faced with the problem of generating the volume and speed data needed to develop mobile source emissions inventories. This problem is compounded by the absence of current time-of-day travel characteristics. Although some of the software vendors that developed programs for the four-step travel demand forecasting process are also developing postprocessors that prepare the 24-hr model output for input to the emissions calculations models, a critical need to be able to respond to clean air agency needs required the development of an immediate response mechanism. The focus is on the application of a travel demand model and an annual traffic count program as the prime ingredients for a process that can be used to convert the 24-hr travel demand model output to peak-hour estimates of travel. The approach is a practical, how-to procedure that enable- the user to estimate volumes and speeds for any hour of the day and for any day of the year, with the ultimate objective of preparing a base year inventory of mobile on-road emissions. Current data sources are evaluated and applied in the process.

The need for peak-hour travel information, particularly estimates of vehicle miles traveled (VMT), has become more pronounced since the Clean Air Act Amendments of 1990 requirement for mobile source emissions inventories to establish conformity. There are several sources of estimated VMT on a transportation network:

- Link-based VMT estimates from regional or statewide traffic counts and roadway segment lengths;
- VMT estimates from fuel sales, vehicle registrations, roadway mileage, population, or a combination of these data;
- Highway Performance Monitoring System estimates for VMT, which are basically derived from traffic count data; and
- Travel demand forecast model estimates of VMT and network travel speeds.

Because most states have some type of travel demand models at the urban, regional, or statewide level, the information generated by these models is a readily available source of future travel estimates that can serve as a base for emissions estimates. The 24-hr models available in these areas usually generate average annual daily traffic (AADT) forecasts that were originally developed to project travel for corridor-level analyses. Some of the software vendors who developed commercial programs to process the traditional four-step travel demand models are

¹Vanasse Hagen Brustlin, Inc., 101 Walnut Street, Watertown, Mass. 02172.

developing postprocessors that prepare 24-hr model output for input to the emissions calculation phase through the application of user-provided factors or program-supplied default values.

The Delaware Department of Transportation (DelDOT) is one of the agencies required to develop a credible and viable base for calculating emission-, for an evening peak 2-hr period. The DelDOT preferred not to use generalized factors or default values in its efforts to derive emission,, estimates. The needs are further compounded by the requirement that speeds for the travel condition,, during the peak period also must be available. In its effort to prepare accurate and reliable motor vehicle emissions for New Castle County (NCC), DelDOT required a process that would provide an opportunity to examine potential emissions levels under a variety of temporal conditions. Estimates of vehicular travel for daily peak periods are required, with an ability to also provide estimates on a seasonal basis. The resulting VMT estimates for 1990 were used to develop a 1990 mobile on-road emissions inventory for the Department of Natural Resources and Environmental Control, the state's clean air agency.

The approach and procedure used to convert DelDOT's NCC 24-hr travel demand model to volume and speed estimates for a 2-hr evening peak period are described. Because of time constraints the technique had to rely on existing data, be practical, and produce credible results.

APPROACH

Two components of model-generated estimate,, need to be addressed: zone-to-zone travel and intrazonal travel. The approach to converting the zone-to-zone estimates generated by the 24-hr models to a peak period relies basically on the manipulation of the existing procedures to enable the user to factor certain of the trip tables that make up the eventual assignment. The method presumes that a valid and calibrated model is available, which is the case with DelDOT. The NCC model configuration is the traditional four-step generation, distribution, model split, and assignment technique. The method for converting 24-hr model data to the peak period is relatively simple in concept and straightforward in application. The trip tables created by the distribution process for internal-internal (I-I) and external-internal (E-I) movements are factored by values that represent the percentage of travel in NCC during the desired hours. External-external (E-E) travel has two basic components of interest: truck and nontruck. These movements are factored by a similar approach. The key to this process is dependent on the availability of the data required to establish the necessary factors.

Three different approaches were considered for this task:

1. Across-the-board factoring of 24-hr link volumes by one average peak period factor,
2. Application of selective peak-period factors, on the basis of traffic count data to different functional classes of roadway, and
3. Adaptation of the 24-hr travel demand model through the application of peak-hour factors by purpose and type of movement.

Some advantages and disadvantages of each approach are given in Table 1. The comparisons are relative to each other and assume that the 24-hr model assignments are generated by a calibrated model.

DelDOT considered the first two approaches to be too generalized and decided to pursue the travel demand model option. The adaptation of the 24-hr model to peak hour is based on the ability to apply peak-hour factors to the various travel components of the system by movement (I-I, E-I, and E-E) and by purpose (work, shop, other, school, non-home-based, and trucks). Trucking is technically a mode but is treated here as a purpose.

The focus in this effort was on the evening peak period, but it can be applied to morning, midday, or anytime of the day by any day or season of the year. For example if peak period volumes are required for days in August considered to be the "hot" days, this approach is also applicable.

Because a 120-min peak period was identified for the emissions calculations, the procedure was applied to two evening peak hours (in this case 4:00 to 5:00 and 5:00 to 6:00 p.m.) separately. This approach is necessary because in the assignment process hourly capacities are used for restraint values. The emissions calculations are performed for each of the peak hours and are summed to obtain a value that represents a peak period.

ZONE-TO-ZONE VMT ESTIMATES

Procedure

If a 24-hr model has been run for the specific year for which emissions calculations are required, the basic components are available and the procedure for the peak-hour processing can be initiated.

The initial steps are concerned with the usual procedures for building a binary network, skimming trees, and updating the skim trees with intrazonal and terminal times. The key difference for peak-hour model processing is that the highway link data records must contain roadway capacities that are expressed in hourly capacities rather than 24-hr capacities. The NCC procedures create

trip tables by purpose for the I-I and E-I movements for 24-hr travel as presented in Table 2.

The I-I person trips are subjected to the model-split process to generate two trip tables: one for transit person trips and one for nontransit person trips. Vehicle occupancy factors, by purpose, are applied to the nontransit person trip table to create a vehicle trip table for the I-I movements. At this point in the process the trip tables represent vehicular AADT volumes and are ready for conversion to peak-hour volumes. Special treatment was applied to certain purposes and movements as shown in Figure 1 and described below.

Convert Work Trips to Peak Hour

The I-I and E-I home-based work (HBW) trip tables are treated differently from those for nonwork purposes because of the peaking characteristics. Work trips tend to be concentrated in the morning and evening. The structure of the trip generation and distribution models creates trip tables that are essentially unbalanced in their original form when the trip interchange pairs represent productions and attractions (P&A). A subsequent step in the process transposes the trip tables into an origin and destination (O-D) format that represents a balanced trip table. The primary difference between the P&A and O-D matrices is the recognition of the zone of

residence. Thus the P&A table is skewed to an A.M., or morning, peak. To appropriately represent an evening peak the P&A trip table is transposed, or flipped, to skew the trip interchange pairs to the evening peak. This treatment of the I-I and E-I HBW trip tables is considered to more accurately portray the directional aspects of work-oriented travel. Factors are applied to convert the work purposes to the peak hour.

TABLE 1 Advantages and Disadvantages to Approaches to Converting 24-hr Model Data to Peak Period

Approach	Advantages	Disadvantages
Across-Board Factoring	<ul style="list-style-type: none"> ● Ease of application 	<ul style="list-style-type: none"> ● Too general ● Does not generate speed data ● Questionable reliability at link level
Link Factoring	<ul style="list-style-type: none"> ● Relatively Simple application 	<ul style="list-style-type: none"> ● Estimates are average and not focused on peaks ● Does not generate speed data
Travel Demand Mode Adaptation	<ul style="list-style-type: none"> ● Refines estimates by considering trip purpose ● Ability to separate truck ● V/C ratios are based on peak hour volumes and capacities ● Provides speed data associated with time period ● Directional splits can be obtained 	<ul style="list-style-type: none"> ● More complex procedure ● Requires detailed traffic count data ● Requires hourly data by purpose

TABLE 2 NCC Trip Purpose Categories

<u>Movement</u>	<u>Purpose</u>	
	<u>Person Trips</u>	<u>Vehicle Trips</u>
I-I	HB Work HB Shop HB Other NHB	Truck/Taxi School
E-I		Trucks Work Shop Other

Convert Nonwork Trips to Peak Hour

Following a logic similar to that applied to work trips, the I-I and E-I nonwork trip components of daily travel have peaking characteristics that are unlike work trip components, and therefore the directionality is not as critical. Home-based shopping, other, and school trips generally have patterns that occur throughout the nonpeak hours and can be appropriately described in the balanced O-D format. The factors for these purposes are applied to the nonwork trip tables generated by the distribution model and balanced by normal procedures.

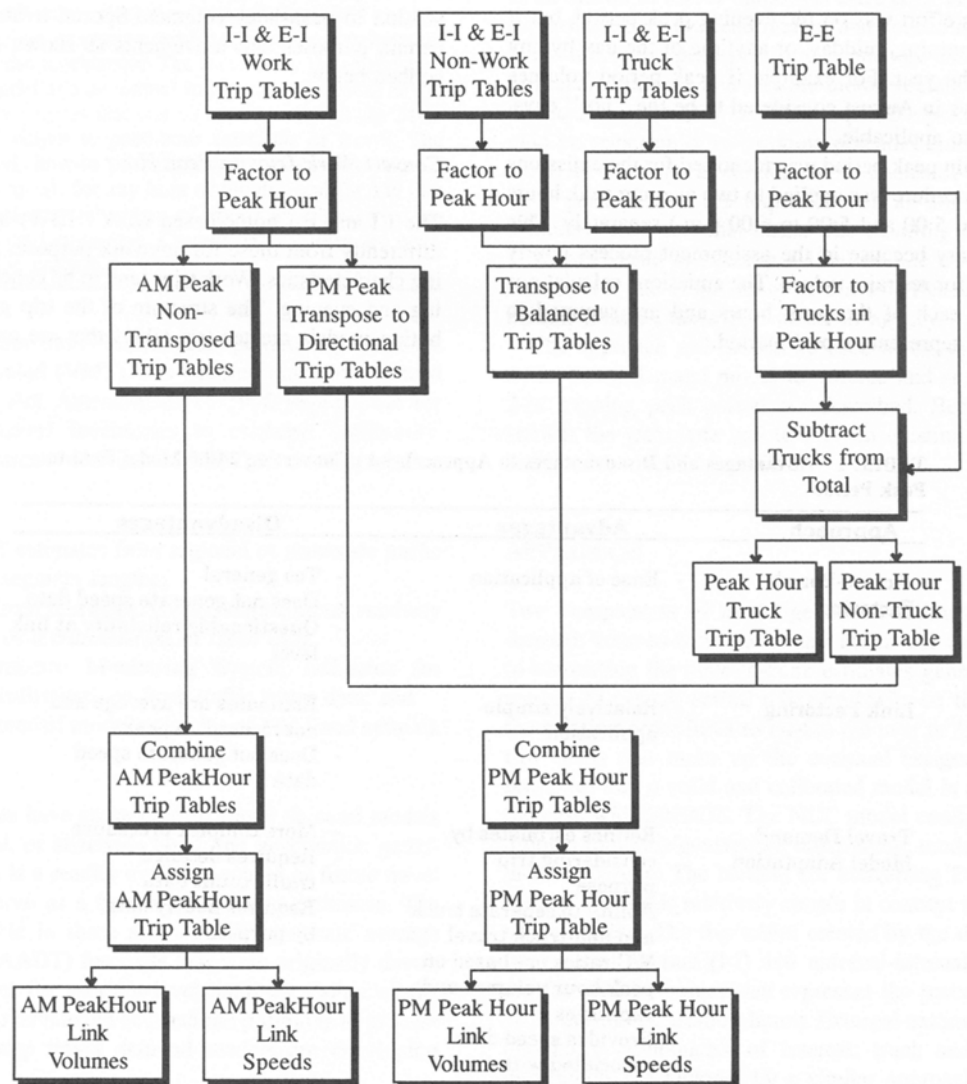


FIGURE 1 Peak-hour conversion process.

Convert Truck Trips to Peak Hour

Two components of truck travel are addressed separately:

- **I-I Trucks.** The movement of internal truck traffic is consistent with the concept of balanced morning and evening peak travel. The factors developed for truck travel described earlier are applied to the trip table transposed into the balanced O-D format.

- E-I Trucks. The movement of trucks into and out of the study area is also considered to be balanced. The key difference is in the factor to be applied to represent the evening peak.

Convert External Travel to Peak Hour

The final component of the travel demand model that requires conversion is the E-E component, or through travel. Traffic count data are available for each external station in the NCC modeling system. Factors were developed for each external to establish two characteristics:

- Peak hour as a percentage of AADT.
- Trucks as a percentage of peak hour.

The factors were applied by using the Fratar technique to model the external stations and establish peak-hour values. The second step was to create a trip table for trucks by applying the truck percentages to the peak-hour trip table, also through the use of Fratar. Finally the peak-hour truck trip table is subtracted from the total-peak-hour trip table to create a peak-hour nontruck trip table. The separation of trucks at this point allows the user to combine the I-I, E-I, and E-E truck trip tables if a special analysis is required or desired for emissions calculations.

Peak-Hour Factors

To make hourly and seasonal estimates possible, significant traffic count data are needed in addition to the 24-hr travel demand estimates. The technique relies on the availability of hourly counts from permanent count stations and classification data to establish hourly volumes as a percentage of daily travel. The classification counts provide estimates of truck and non-truck data. The following kinds of data are available from DelDOT for NCC:

- Permanent Count Stations. There are 20 stations in NCC at which 24-hr counts are recorded hourly for 365 days each year. The data from each of these locations were arrayed to calculate hourly percentages. Directionality was also maintained in the process to identify directional splits. The resulting information formed the base for identifying the evening peak hours and the percentage of the total 24-hr count of each one.
- Traffic Count Stations. Shorter-term counts are taken at various locations throughout NCC and are expanded to an AADT volume. The report in which these counts are summarized by maintenance road section also includes peak-hour and truck percentage information.

The peaking characteristics of traffic on a regional basis tend to remain consistent over time, and the peak-hour data used in this process can be assumed to remain valid for forecast years. However the peak-hour percentages should be evaluated for applicability to forecast years, particularly if major land use changes are anticipated to occur, as evidenced by the allocation of population or employment growth in certain traffic analysis zones.

Peak-hour truck travel percentages for I-I and E-I movements were borrowed from an FHWA report (1). Of the data provided, Louisville, Ky., appeared to be an area most comparable to NCC. Figure 2 presents the percentages of internal and external truck travel by hour of day. Since the application of these data DelDOT has translated its traffic count data into a geographic

information system, which will expedite access to time-of-day data for trucks and nontrucks for future efforts.

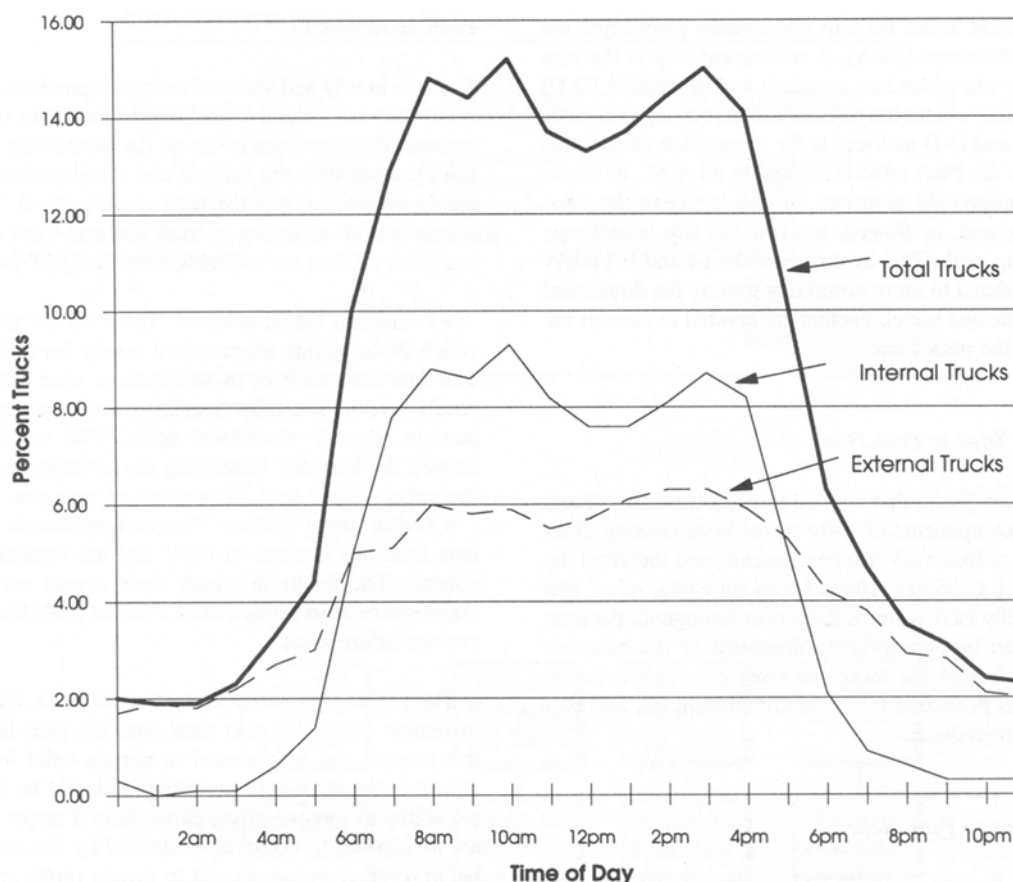


FIGURE 2 Hourly distribution of truck travel (I).

Information regarding the percentage of travel by purpose during the peak hour is also required for nontruck travel. Because the travel demand models were developed in the mid-1960s in most states, the source data from the household surveys were aggregated to zonal averages and the original detail was lost. Therefore a reconstruction of travel characteristics was not possible. This situation required that a more current source of time-of-day travel data by purpose be identified, evaluated, and applied. One reliable source of current information is the Nationwide Personal Transportation Survey (NPTS). The percentage of travel by purpose [work, shop, other, non-home-based (NHB)] can be estimated from these data. For this effort a special tabulation was generated to create a matrix of trips by purpose and hour of day. Because of the survey's sample size the information could not be focused on Delaware specifically. The most statistically reliable level available was for the South Atlantic Region, which covers the area along the East Coast from Delaware to Florida.

To convert the 24-hr NCC travel demand model to represent peak-hour travel, a series of factors had to be developed. The nontruck I-I and E-I purposes were factored with values derived from NPTS. These distributions are represented graphically in Figure 3. It would have been more desirable to have such information in smaller time increments, such as 15-min slices, to more accurately evaluate peak-period spreads. Although the data might indicate a peak hour from 4:00 to 5:00 p.m., it might actually be 4:15 to 5:15 p.m. Because the DelDOT traffic counts and the NPTS data are reported in hourly increments, an inspection of the peak-hour percentages for the 20 permanent count stations in NCC suggested evening peak hours of 4:00 to 5:00 and 5:00 to 6:00 p.m. A similar review of the NPTS data also showed evening peaks of 4:00 to 5:00 and 5:00 to 6:00 p.m. Therefore the percent breakdowns for trip purposes according to NPTS were applied to the NCC data as shown in Table 3.

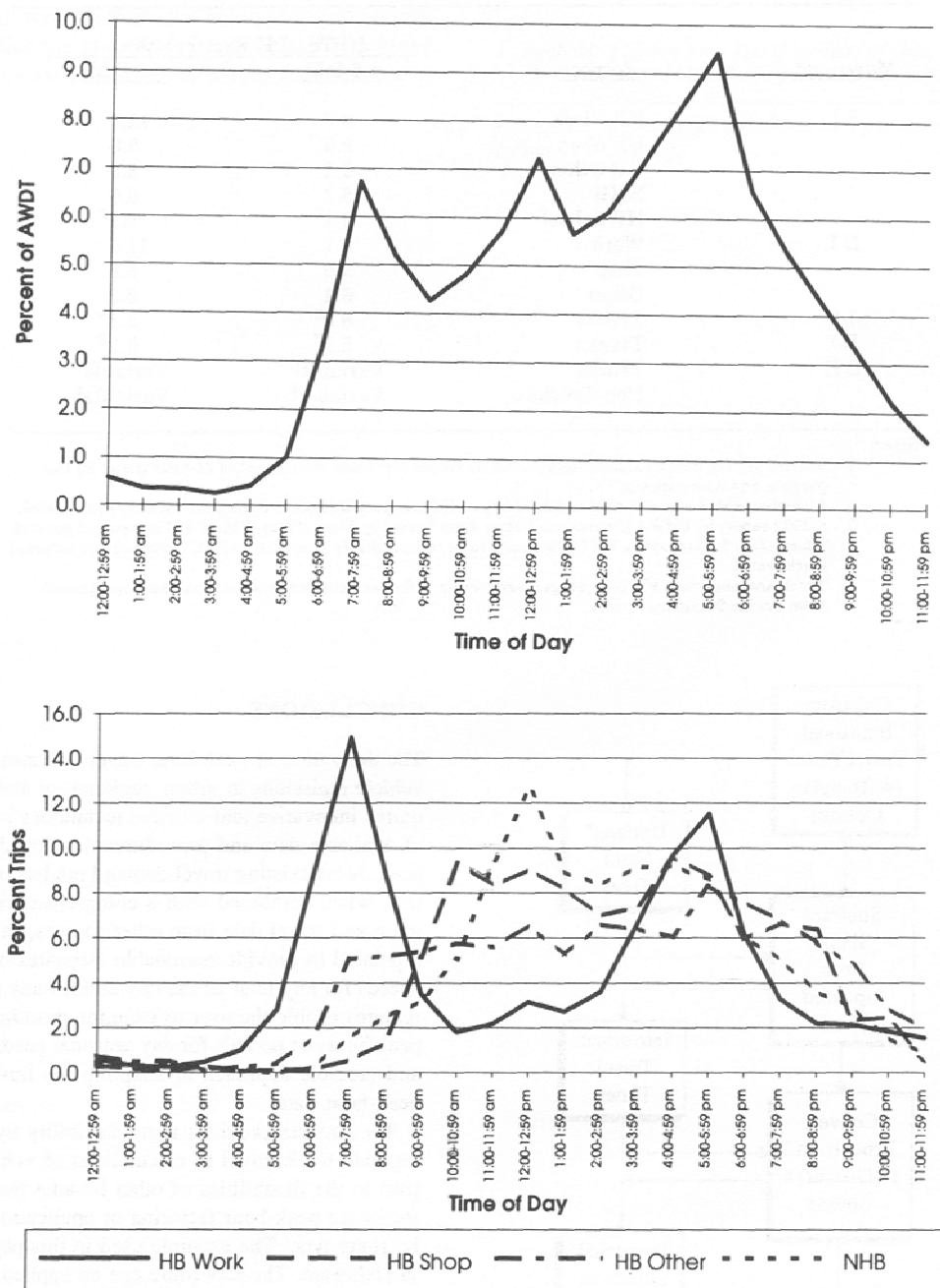


FIGURE 3 NPTS data for total trips (top) and trips by purpose (bottom).

TABLE 3 Peak-Hour Factors (1)

<u>Movement</u>	<u>Purpose</u>	<u>NPTS - PH Factors¹</u>	
		<u>4-5 P.M.</u>	<u>5-6 P.M.</u>
I-I	HB Work	9.7	11.5
	HB Shop	9.6	8.8
	HB Other	6.1	8.3
	NHB	9.2	8.6
	HB School	0.1 ²	0.1 ²
E-I	Work	9.7	11.5
	Shop	9.6	8.8
	Other	6.1	8.3
	Trucks	8.2 ³	5.1 ³
E-I	Trucks	5.9 ³	5.3 ³
E-E	Trucks	Variable ⁴	Variable ⁴
	Non-Trucks	Variable ⁴	Variable ⁴

Notes:

1. Source: NPTS, South Atlantic Region. These values represent the percent of 24-hour travel for the purpose and hours shown.
2. A nominal value was assigned to school trips (HBSC) expected to occur during the evening peak period.
3. A 1972 report by FI-RWA (Analysis of Urban Area Travel by Time of Day, FM-11-7519) reported general information for Louisville, KY (determined to be comparable in character to NCC) internal and external truck travel.
4. Truck and non-truck pH percentages were obtained for each external station from the Department's 1990 Traffic Summary Report.

The identification of hourly travel percentages for external stations was accomplished by processing of permanent traffic count data in NCC. Data from selected stations were summarized to establish the 24-hr increments as a percentage of the total. When regional travel demand models are developed, the travel inventory usually records data for travel on nonholiday weekdays, and the resulting trip generation relationships represent average weekday traffic (AWDT). To obtain an estimate of travel that was as accurate as possible for application to emissions modeling, AWDT counts were obtained by deleting weekend days and major holidays from the permanent count station data. Peak-hour percentage factors were developed from the traffic count data for each of the external stations.

INTRAZONAL VMT ESTIMATES

Because intrazonal trips are not loaded onto the network, estimate of travel, or in this case VMT, are underestimated. Furthermore for emissions calculations intrazonal trips are usually made at speeds lower than the speeds that trips on the rest of the system are made, thereby probably creating a relatively greater volume of pollutants.

The options for estimating intrazonal VMT are somewhat limited. Because the effort to derive link VMT and speeds had a practical orientation, available resources were used. NCC tree building and skimming were accomplished with DelDOT's software of choice, which has an option that allows the user to estimate intrazonal travel times by averaging the travel times to adjacent zones. This approach generally recognizes the size of the zone and results in a reasonable approximation of the intrazonal travel times.

For the NCC system an average intrazonal travel speed of 15 mph was used. Given the time value of intrazonal travel and an average intrazonal speed the distance was easily calculated. When the intrazonal travel distances are calculated they are arrayed as the diagonal cells of a to-from matrix. Likewise the intrazonal volumes in a trip table are represented in a similar diagonal. To obtain a total intrazonal VMT estimate the values in the diagonals of the distance matrix and the factored trip tables for each purpose are multiplied and summed. Figure 4 graphically represents the process for estimating intrazonal VMT.

ASSIGNMENT PROCEDURE

The final step in estimating peak-hour travel is to combine each of the factored purposes and movements into one trip table for use in the assignment process. For NCC the equilibrium technique was used, with the assigned volumes restrained by the hourly capacities represented on each of the network links. Speeds were also saved to provide the user with the volumes and speeds for each link in the network that are received to calculate emissions on a link-by-link basis.

Because emissions estimates were required for a peak 120-min period and because the assignment technique uses capacity as a restraint, two peak-period hours were selected and separate assignments were executed. The emissions were then calculated individually for each hour and were combined for the total peak period.

At the time of the effort described here the base traffic count data in Delaware were recorded in hourly increments, and the hour was the smallest time unit that could be applied. Some data bases might be available in 15-min increments, thereby providing the user with more flexibility in defining the peak period.

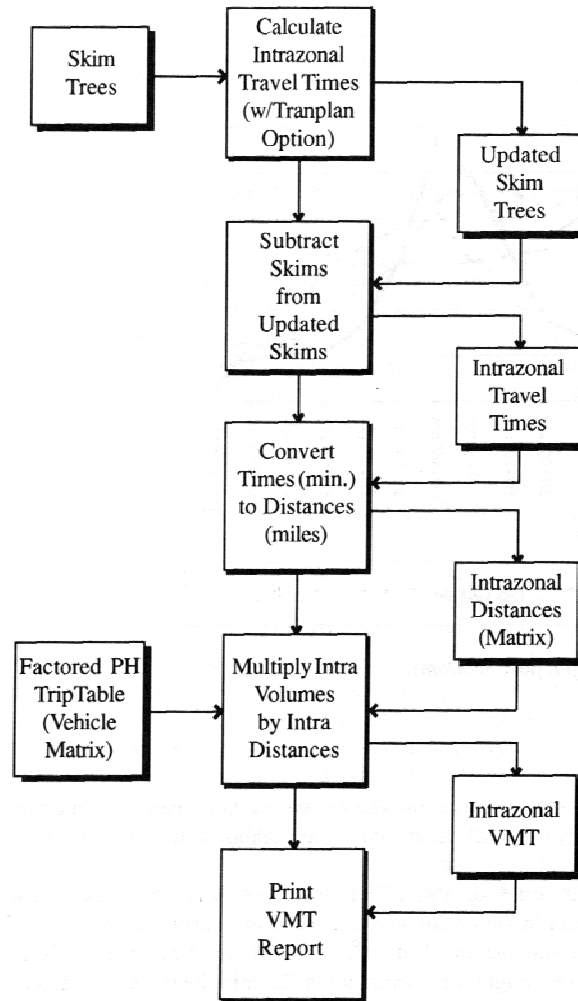


FIGURE 4 Intrazonal VMT estimate process.

CONCLUSIONS

The derivation of peak-hour travel volumes for the calculation emissions in urban, regional, or statewide study areas requires innovative and creative techniques because of the absence of available data and procedures developed for that specific purpose. Most existing travel demand models offer a viable data base that, when combined with a comprehensive traffic counting program and travel data from other sources, can be factored and manipulated to provide reasonable estimates of hourly volumes and speeds for any hour of the day and for any day of the year. These features enable the user to examine morning, midday, or evening peak hours or periods for any seasonal needs and offers a practical and credible approach to adapting the travel demand models to peak-hour data.

The process provides more flexibility by allowing the user to separate truck travel in calculations of vehicle emissions in contrast to the flexibilities of other broader methods, such as across-the-

board peak-hour factoring or application of peak-hour factors by route type. The example used in this paper is specific to NCC in Delaware. The procedure can be applied for most model structures to represent the variety of temporal conditions required to prepare a mobile on-road emissions inventory. Estimates of emissions, and particularly reductions in emissions owing to improved transportation system components for future transportation system scenarios, will depend on application of the same procedures and factors described above. As with most modeling techniques base year relationships are assumed to carry forward to the future. With the types of data used to factor the 24-hr models by the procedure described here, there is an opportunity to review trends over time and make adjustments as necessary.

ACKNOWLEDGMENTS

The authors wish to recognize the assistance provided by members of the Delaware Department of Transportation's Intergovernmental coordination Section, namely, Anthony D. Peer and Robert Shiuh. Peer and Shiuh were instrumental in supply in the necessary base files on which the procedure described here was based and made valuable review contributions during development of the procedure.

REFERENCE

1. *Analysis of Urban Area Travel by Time of Day*, Table 1-6. Report FM-11-7519. FHWA, U.S. Department of Transportation, 1972.

Publication of this paper sponsored by Committee on Passenger Travel Demand Forecasting.

Shopping Trip Chains: Current Patterns and Changes Since 1970

HYUNGHIN KIM, ASHISH SEN, SIIM SÖÖT, AND ED CHRISTOPHER¹

National demographic patterns are changing. In absolute terms household and automobile growth exceeds population growth, resulting in an increase in the number of trips and traffic congestion. The evidence from the Chicago region, however, suggests that the number of trips per capita has not changed in 20 years, trip chains per capita are declining, travel per households has declined, and perhaps most surprisingly shopping trips per capital have declined noticeably. However, through increasing trip-chain complexity, more out-of-home destinations are reached with a constant number of trips, indicating a higher degree of trip mobility. Trip making is becoming more efficient, and the time spent shopping is not increasing even with fewer per capita shopping trips. Although many of these trips are conducted during the peak and add to congestion, since they are chained with the work trip, moving these trips to the off peak may increase vehicle miles of travel (VMT). Moreover, relatively few trip chains follow the minimum path, thereby, adding to VMT. These conclusions are drawn from 1970 and 1990 household travel data collected by the Chicago Area Transportation Study for DuPage County, a fast-growing area west of Chicago. The authors encourage others to examine the temporal changes in travel behavior in their locales.

During the last several decades the transportation community began to change its focus of travel analysis from individual trips to trip chains. This change acknowledges the importance of multipurpose trip making. Concurrently there has been a proliferation of models addressing trip chaining (1-4). Although considerable strides were made in modeling this behavior, the demographics and the demand for transportation services have changed, affecting these model constructs. The number of jobs has grown disproportionately to population growth as women have entered the labor force in large numbers. Rapid job formation encouraged the automobile population to grow rapidly, which contributed to a decline first in transit use and subsequently in carpooling. All of these factors together with the new questions being asked regarding the environmental effects of travel suggest a need to review existing travel demand models.

The purpose of the study described here is to examine some of the current chaining characteristics as they describe shopping trips and more generally, to identify some of the changes in trip chaining behavior since 1970. Although the emphasis is on current shopping trips, other trips are also examined, particularly in contrasting 1970 and 1990 data. The study concludes that as the number of work trips per capita has increased, the number of shopping trips has declined, despite the implicit increase in income stemming from job growth. Moreover, the time spent shopping is not increasing, despite the declining frequency of these trips. As has been

¹ H. Kim, A. Sen, and S. Sööt, Urban Transportation Center, MC 357, 1033 West Van Buren Street, Suite 700, University of Illinois at Chicago, Chicago, Ill. 60607-2919. E. Christopher, Chicago Area Transportation Study, 300 West Adams Street, Chicago, Ill. 60606.

asserted by many (5-7), changes in household structure have resulted in modifications in trip-making behavior, but as discovered here, because trips are increasingly linked together, there has been little change in the number of trips per person. Nevertheless the number of out-of-home stops has increased, made possible by increasing chain complexity. Mobility seems not to suffer since more destinations are reached with fewer trips and chains.

Lastly in the study of shopping trips and chains it is inevitable that one becomes involved with other trip purposes. Shopping chains frequently include many nonshopping stops.

BACKGROUND: TRIP CHAINING

Definitions

Several terms need to be defined or clarified before proceeding. First, a *chain* is defined as a series of trips that begin and end at home. A *trip* is the movement or link from one stop to another. A *shopping trip* is a trip in which shopping is the purpose at the destination. Second, chains can consist of any number of stops and may have any combination of purposes. Shopping chains, then, may have numerous nonshopping stops but at least one stop must be to shop. Third, home-to-shop-to-home is a simple shop-ping trip chain; complex chains have more than two out-of-home stops. Home-to-work-to-bank-to-shop-to-home is an example of a complex shopping trip chain. Because of the process of linking stops together and the definitions, shopping stops constitute a minority of nonhome stops in complex chains.

Previous Studies

A wide variety of approaches has been developed, beginning largely from a Markovian base (8,9). Subsequently advances were made in formulating the theoretical basis for trip chaining (10,11), and a method has also been provided to estimate the amount of trip chaining (12). Many of these papers include extensive discussions regarding previous work (13,14), including trip chaining as it pertains to pedestrian travel (15); therefore, it is not necessary to restate these developments.

Two studies merit attention, however. These empirical studies have examined trip chaining with data collected in the last ten years. Strathman (7) examined data collected in Portland, Oregon, and addressed the degree to which nonwork trips were chained to work trips. Compared with DuPage County, Illinois, the Portland study found a lower propensity to conduct complex chains: 24 percent of all trip chains versus 37 percent in DuPage County. Simple chains to work and to shop, however, were found in similar proportions. Home-shop-home accounted for 9.6 percent of all trip chains in Portland and 9.0 percent in DuPage County. The respective figures for simple work trip chains are 25.1 and 23.0 percent.

The recent examination of travel data collected in the Seattle region between 1986 and 1989 shows the greatest amount of chaining by women in suburb-to-suburb trips and the fewest complex chains by men from the suburbs to the city (16). It also identified a high frequency of trip chaining by women during the midday. Given the general propensity to trip chain, it was

concluded that transit could not well serve complex chains and that transit potential was consequently negatively affected by this phenomenon.

DATA AND STUDY AREA

Household and Survey Data

The data used for the present analysis were extracted from two different travel data bases used by the Chicago Area Transportation Study (CATS), the metropolitan planning organization for northeastern Illinois. The first set of data was selected from a CATS 1970 home interview survey that contained a 0.8 percent sample, or 17,385 households. For just over 20 years it was this data base that was used by CATS in most of its travel forecasting and planning work. A total of 1,110 households represent DuPage County, the area examined in the study.

TABLE 1 DuPage County Data Bases, 1970 and 1990

Description	Number of Households		Number of Trip Chains		Number of Trips	
	Universe	Sample*	Total	Shopping	Total	shopping
1970 Survey						
Regionwide	2,183	17,385*	6,798	1,676	16,757	2,101
DuPage County	142	1,110*	619	171	1,518	210
1990 Survey						
Regionwide	2,773	19,313*	NA	NA	NA	NA
DuPage County	279	5,098*	792	184	2,178	226
Percent Increase in DuPage County	96%	—	28%	8%	43%	8%

* All numbers in thousands except for sample size

Source: Chicago Area Transportation Study (CATS) 1970 Home Interview data base, 1990 Household Travel Survey results and Preliminary factored results for DuPage County 1990 data.

The second data source was the recent (1988 to 1991) CATS Household Travel Survey, which consisted of a 0.7 percent sample of households regionwide and a 1.7 percent sample in DuPage County. For the region this data base contained information from 19,313 households, of which 5,098 were in DuPage County. Table 1 presents the size of the data bases with a focus on the expanded number of trips and trip chains in DuPage County.

Both surveys and their resultant data bases have been well documented, and each carries a wealth of information (17-19). For the 1990 data only the DuPage County portion has been factored and adjusted. Consequently these data are preliminary. However the final data are expected to closely match the preliminary set.

TABLE 2 DuPage County Population Characteristics, 1970 and 1990

	1970	1990	% Change
Population	491,882	781,666	58.9
No. of HH.	142,408	279,344	96.2
No. of Person per HH.	3.41	2.76	
No. of Workers	199,352	425,284	113.3

DuPage County

For the decade of the 1980s DuPage County demonstrated one of the largest absolute population gains outside of the Sun Belt states. In 1970 the county had 491,882 residents and 199,352 workers (Table 2). By 1990 the population had swelled to 781,666 people, 425,284 of whom were employed. In terms of jobs the increases were staggering. In 1970 there were approximately 115,200 jobs, which grew to 528,444 in 1990, an increase of 359 percent (20). DuPage County is representative of a fast-growing suburban community.

TABLE 3 Mean Trip Lengths of Shopping Trips by Trip Chain Sizes

Chain Size	Link 1	Link 2	Link 3	Link 4	Link 5	Link 6
2	2.51	2.51*				
3	2.76	3.75	2.73*			
4	3.09	3.41	3.63	3.71*		
5	2.81	2.65	3.62	3.81	4.00*	
6	5.44	3.55	2.97	3.34	4.73	4.21*

Only trip chain sizes 2 to 6 included in this table

* Returning home trip; not a shopping trip

Unit = zone centroid to centroid airline distances in miles

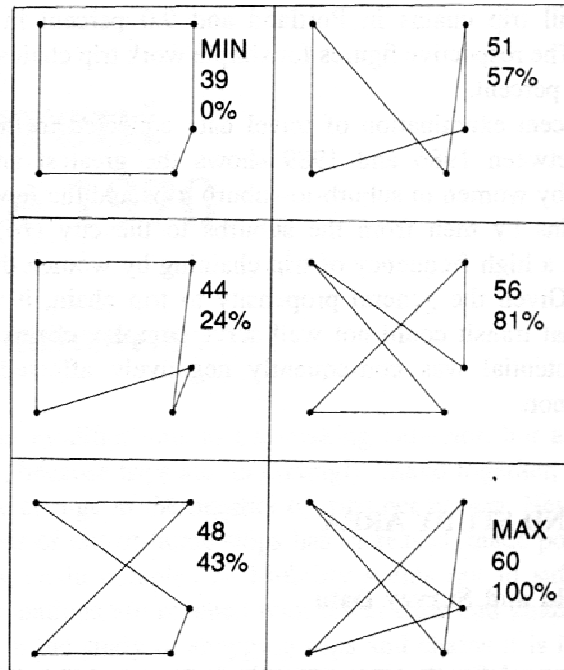


FIGURE 1 Six path choices with travel distances and percentage of range from minimum to maximum.

SUMMARY OF 1990 SHOPPING TRIPS AND CHAINS

A review of the number of shopping trips and shopping chains in DuPage County reveals that trip chains are for a variety of purposes and that only a small proportion of chains are for a single purpose. There are a total of 619,170 trips in the shopping trip chains, but only 36.6 percent (226,399) of these are shopping trips. Further examination reveals that 58.7 percent of the shopping trips are made in conjunction with other purposes. Shopping *chains* include nonshopping activities, whereas shopping *trips* refer only to trips with shopping at the destination.

TABLE 4 Distance Minimization in Path Chosen by Shoppers by Chain Sizes

Frequency of Chosen Paths (in percent)						
Chain size	Minimum	0 – 20%	20 – 60%	60 – 100%	Maximum	Number of Chains
4	47.3	21.3	5.8	3.8	21.7	520
5	16.7	35.5	21.8	16.4	9.5	293
6	8.8	42.2	33.3	12.9	2.7	147

Mean Values of Range between Maximum and Minimum Distance

Chain size	Minimum	0 – 20%	20 – 60%	60 – 100%	Maximum	Number of Chains
4	3.7	6.7	1.7	1.7	2.6	520
5	7.3	12.7	6.0	4.4	4.0	293
6	11.2	20.0	9.8	8.3	4.2	147

Trip and Chain Lengths

At least two aspects of trip and chain lengths are of importance: (a) the number of miles and (b) the degree of distance minimization in the trip chains. The trip lengths were derived by determining the quarter-square-mile origin and destination zones and then computing the airline distance between the zone centroids. Distance minimization pertains to the sequence of stops and how closely this route comes to the minimum path through the stops.

Trip Lengths

An examination of trip lengths by link and chain size reveals important patterns. The first link was only 2.51 mi. in simple shopping day in the Chicago-area market. It is likely that these two counteracting effects may balance, and if they do not, there may be a slight bias toward more shopping being recorded in the 1990 data. The authors found, however, the since per capita shopping trips have declined, there is little evidence of a proshopping bias. As a whole care must be exercised in drawing precise comparisons from data such as these.

TABLE 5 Path Choice Behavior of Shopping Trip Chains by Chosen Path Category (chain size of 4)

Description		Mini- mum	0 - 20%	20 - 60%	60 - 100%	Maxi- mum
Number of Trip Chain		246	111	30	20	113
Number of Work Trips		74	32	10	9	75
Activity Time (minute)		237	252	274	349	324
Mean Distance of Trip Chains		13.9	18.3	14.4	21.9	18.9
Mean Distance between Min and Max		3.7	6.7	1.7	1.7	2.7
Number of Trip-chains by Distance Traveled (miles)	0 - 5	34	10	2	2	11
	5 - 10	79	27	9	4	26
	10 - 20	89	37	15	5	38
	20 - 30	22	18	2	3	19
	30 +	22	19	2	6	19

Socioeconomic Characteristics of Individuals

Mean Age		46.2	46.5	43.3	47.1	43.1
Number of Trip-chains by Gender	Male	62	37	8	7	38
	Female	184	74	22	13	75
Number of Trip-chains by Employment status	Full Time	99	52	14	8	58
	Part Time	53	15	5	4	17
	Homemaker	69	28	9	3	32
	Student	24	9	4	1	16
	Retired	48	26	2	4	14

Changes in Numbers of Trips and Trip Chains

With the population expansion there has been a growth in the number of daily trips, from 1.5 million to 2.2 million, and an increase from 619,000 to 792,000 trip chains (Table 1). Nevertheless this represents a decline in several categories: trip chains per household, trips chains per person, and trips per household (Table 7). The only rate that remained stable was trips per capita. On the surface this may seem surprising but it is in keeping with (a) the trends displayed by the Nationwide Personal Transportation Survey (NPTS) (22) and (b) the expectation that trips would be bundled into chains as time constraints mount. The 1983 NPTS shows a decline in the number of trips per capita from the 1977 survey, but the 1990 NPTS figure is approximately 5 percent higher than that in 1977. Suffice it to say that given large increases in the automobile population, there has been remarkably little change in the number of trips per capita on the basis of both NPTS and DuPage County data. The declines in the other three rates may be attributed to demographics. The trips and trip chains per household rates are declining because of smaller household sizes.

The decline in per capita trip chains is plausible even with increasing mobility. Figure 2 illustrates two hypothetical households. Household A completes three simple chains, visiting three out-of-home destinations. Household B, however, completes only one chain with five trips but visits one more out-of-home site. It is therefore possible to visit more sites with fewer chains

and with fewer trips. This has occurred in DuPage County. Although the number of trips per person has declined modestly, from 4.3 to 4.2 trips per day per person, there has been a 13 percent increase in the number of out-of-home sites visited between 1970 and 1990. This marks a Significant modification in which individuals can reach more destinations with less travel; that is, they are more "mobile" but travel less.

TABLE 6 Frequency of Sample Shopping Trip Chains

Chain Size	Chain Type	Percent
4	H-S-X-S-H	4.8
	H-S-S-X-H	6.0
	H-S-S-S-H	6.5
	H-S-X-X-H	9.2
	H-X-S-S-H	14.4
	H-X-S-X-H	26.9
	H-X-X-S-H	32.1
5	H-S-S-S-S-H	2.0
	H-S-X-X-X-H	6.1
	H-X-X-S-X-H	19.1
	H-X-X-X-S-H	31.7
6	H-X-X-X-X-S-H	21.0

H = home, S = shop, H \neq X \neq S

Changes in Shopping Chains

With increased trip chaining each excursion from the home includes more destinations, and the instances of travelers conducting a trip for only one purpose on a chain are decreasing. The authors are therefore rather liberal in their definition, which includes all chains that have at least one shopping destination; many shopping chains include more nonshopping stops than shopping stops.

TABLE 7 Comparisons of 1970 and 1990 DuPage County Daily Travel Data

Variable	1970	1990
No. of Trips*	1.5	2.2
No. of Trips/Person	4.3	4.2
No. of Trips/Household	10.3	9.1
No. of Chains/Person	1.8	1.3
No. of Chains/Household	4.3	2.8
No. of Trips/Chain**	2.3	2.9

* Number in millions

** Change mode is excluded

Numbers of Shopping Chains and Shopping Trips

As with all travel shopping chains and shopping trips have increased since 1970 but not always in proportion to population. There has, however, been a dramatic decline (30.6 percent) in the number of simple shopping chains (Table 8). Other simple chains have increased in number such as trips to work, but the practice of going to a store and returning has declined even with the large population increase.

TABLE 8 Number of Trip Chains per Day in DuPage County, 1970 and 1990

Chain Size	All Chains			Shopping Chains		
	1970	1990	Percent Change	1970	1990	Percent Change
2	481,582	499,216	+3.7	102,986	71,484	-30.6
3	85,200	137,905	+61.9	47,573	53,381	+12.2
4	32,828	83,910	+155.6	12,815	28,122	+119.4
5	10,491	33,857	+222.7	4,443	15,349	+245.5
6+	9,299	37,168	+299.7	3,453	16,653	+382.3
Total	619,400	792,056	+27.9	171,270	184,989	+8.0

To compensate there have been significant increases in the number of complex shopping chains, most noticeably in chains with four and more destinations, all of which have more than doubled in number. Still an increase of only 8 percent in the number of shopping trips is unexpected. Increasing from 210,000 to 226,000, the rise does not reflect the 60 percent increase in population, let alone the increasing disposable income brought about by increased participation in the labor force.

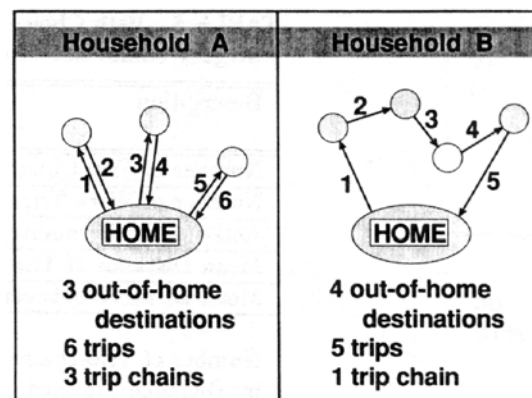


FIGURE 2 Two hypothetical trip patterns.

Duration of Shopping

The number of shopping trips per capita has declined even with the proliferation of shopping centers and scattered retailing sites. It would seem reasonable that to accommodate the needs

met by shopping, the amount of time spent shopping might increase. Mitigating against that is the increase in the number of people in the labor force and the consequent constraints on time to shop. Also as chain complexity increases, the average time spent at each stop tends to decrease.

The 1990 CATS survey shows an average amount of time spent at the shopping destination to be 42 min., down from 49 min. in 1970 (travel time not included; Table 9). An examination of the shopping duration distribution indicates that the greatest change was in the decline in long shopping trips (in excess of 90 min.), which can be attributed to time constraints common to multiple-worker households.

TABLE 9 Frequencies of Shopping by Duration Categories (time spent at each shopping destination; travel not included)

Duration (minutes)	1970 (percent)	1990 (percent)
0 - 14.99	20.6	23.3
15 - 29.99	21.8	25.1
30 - 44.99	17.1	17.8
45 - 59.99	13.9	11.7
60 - 89.99	11.8	13.3
90 +	14.8	8.8
Total	100.0	100.0
Average	49 minutes	42 minutes

It appears that the constraint on people's time was a slightly stronger force, resulting in a decrease in shopping duration. This can be seen in the declining duration at each shopping stop as the trip chain becomes more complex. The amount of time spent shopping in simple chains was 49.8 min., and it declined to 30.8 min. for chains of seven and more links, for a difference of almost 20 min. The increasing tendency over time to trip chain makes shopping a more directed activity and is less of a social or recreational experience, which was more frequently the case in the past. This reduces the fuzzy distinction between shopping and recreation, which is common to some shopping trips and which contributes to poorly defined trip purposes in transportation surveys.

Shopping Trip Lengths

There are also at least two competing forces on shopping trip lengths. First, the increased number of shopping sites throughout DuPage County has brought many more shopping choices closer to places of residence and therefore would contribute to shorter shopping trips. Second, the shortest links in 1990 were in the simple chains, which have declined precipitously since 1970 (Table 3). These links averaged 2.5 mi., whereas most others averaged over 3 mi. and some averaged more than 4 mi. As these simple chains decline the average distance should increase.

The data show that the change in average shopping trip distance has stayed stable, rising only from 3.08 mi. in 1970 to 3.11 mi. in 1990

(Table 10). The distance distribution has also changed very little. Approximately 60 percent of all shopping trips in both surveys were less than 2.5 mi., and roughly 1 in 20 was more than 10 mi. Indeed the slight increase in trip length can be attributed to the modest decline in short trips (less than 2.5 mi.), many of which were simple chains. It appears that the decline in simple chains was stronger than the effect of increased density of stores, which permits shorter trips.

TABLE 10 Frequencies of Shopping by Distance Categories

Distance (miles)	1970		1990	
	Frequency	Percent	Frequency	Percent
0.0 – 2.49	131,068	62.3	134,594	59.5
2.5 – 4.99	42,335	20.1	52,836	23.3
5.0 – 9.99	27,214	12.9	28,282	12.5
10.0 +	9,786	4.7	10,678	4.7
Total	210,403	100.0	226,399	100.0
Average	3.08 miles		3.11 miles	

CONCLUSIONS AND IMPLICATIONS

Although there are differences between how 1970 home interview and the 1990 Household Travel Survey data were collected and the questions asked, it is possible to identify broad findings. The most significant finding is that trip making appears to be more efficient and shopping as an activity is declining as the number of per capita work trips rises. There is little change in the number of trips per capita, but since there are more complex trip chains, travelers achieve more out-of-home stops with a fixed number of trips. It takes six trips to visit three places if all are simple chains, but if these six trips are linked in one chain, five places can be visited, almost twice the number. Still the increase of only 8 percent for shopping trips was unexpected, given the population increase and the rising incomes through an expanded labor force.

With the decline in the number of simple shopping chains, which are typically short trips, there is the potential for average trip lengths to increase. This tendency is partially but not completely offset by a greater density of stores, which contributes to shorter trips.

Despite the stable shopping trip distances and the small rise in the total number of shopping trips, highways are becoming more congested, and there are few better examples than DuPage County. In DuPage County the gross densities are very low, the population is affluent, vehicle ownership is high, and the populace is modifying travel behavior by stringing trips together into complex chains. This may be a reaction to less actual or perceived leisure or nonwork time, but the consequence is that the total travel in 1990 was more efficient for the individual than in 1970, but it contributed to severe peaking and congestion.

Contemporary travel behavior increasingly links trips for nonwork activity to the work trip, many of which occur during the peak. It may reduce congestion if trips for nonwork activity were rescheduled to other times, but this would be a return to the 1970 pattern, in which a large

number of simple chains characterized household travel. This could add to vehicle miles of travel (VMT) and may not be desirable unless work travel occurred during the off peak or the shoulder of the peak, as would be the case with staggered work hours. These travel patterns also have implications for cold starts, which would also likely increase if chaining declined. Additional work is encouraged to ascertain the merits of peak-hour trip chaining and the trade-off between reduced congestion and increased VMT.

There are also clear implications for trip distribution modeling. These models need to consider more closely the origin and destination of the trip since many trips are now made without the home at either end. yet the location of the home is undoubtedly important in selecting nonhome destinations.

This analysis underscore, the fact that there are considerable archives of travel data, perhaps more than can be analyzed. Still there are many unverified conjectures about how travel has changed. The authors encourage other work to examine these changes and explore ways that this work can be used to improve the transportation modeling and planning process.

REFERENCES

1. Goulisa, K. G., and R. Kitamura. Recursive Model System for Trip Generation and Trip Chaining. In *Transportation Research Record 1236*, TRB, National Research Council, Washington, D.C., 1989, pp. 59-66.
2. Lerman, S.R. The Use of Disaggregate Choice Models in Semi-Markov Process Models of Trip Chaining Behavior. *Transportation Science*, Vol. 13, 1979, pp. 273-291.
3. Adler, T., and M. Ben-Akiva. A Theoretical and Empirical Model of Trip Chaining Behavior. *Transportation Research B.*, Vol. 13, 1979, pp. 243-257.
4. Timmermans. H.. X. Van der Hagen, and A. Borgers. Transportation Systems. Retail Environments and Pedestrian Trip Chaining Behavior: Modeling Issues and Applications. *Transportation Research B*, Vol. 26.1992, pp. 45-59.
5. Oster. C. *Second Role of the Work Trip-Visiting Non-Work Destinations*. In *Transportations Research Record 728*, TRB, National Research Council, Washington, D.C., 1979, pp. 79-82.
6. Prevedouros, P., and J. Schofer. Suburban Transport Behavior as a Factor in Congestion. In *Transportation Research Record 1237*, TRB, National Research Council, Washington, D.C., 1990, pp. 47-58.
7. Strathman, J. G., K. J. Dueker, and J.S. Davis. Effects of Travel Conditions and Household Structure on Trip Chaining. Presented at 72nd Annual Meeting of the Transportation Research Board, Washington, D.C., January 1993.
8. Horton. F. E. and W E. Wigner. A Markovian Analysis of Urban Travel Behavior: Pattern Response by Socioeconomic Occupational Group.. In *Highway Research Record 283*, HRB, National Research Council. Washington, D.C., 1969, pp. 19-29.

9. Hanson. S. Urban Travel Linkage: A Review. In *Behavioral Travel Modeling* (D. Hensher and P. Stopher, eds.), Croom Helm, London, 1979, pp. 81-100.
10. Bacon, R. *Consumer Spatial Behavior*. Clarendon Press, Oxford, 1984
11. Gosh, A., and S. McLafferty. A Model of Consumer Propensity for Multipurpose Shopping. *Geographical Analysis*, Vol. 16, 1984, pp. 244-249.
12. Goulisa, K. G., R. M. Pendyala, and R. Kitamura. Practical Method for the Estimation of Trip Generation and Trip Chaining. In *Transportation Research Record 1285*, TRB, National Research Council, Washington, D.C., 1990, pp. 47-56.
13. Nishii, K., K. Kondo, and R. Kitamura. Empirical Analysis of Trip Chaining. In *Transportation Research Record 1203*, TRB, National Research Council, Washington, D.C., 1988, pp. 48-59.
14. Thill, J. C., and I. Thomas. Toward Conceptualizing Trip-Chaining Behavior: A Review. *Geographical Analysis*, Vol. 19, 1987, pp. 1-17.
15. Borgers, A., and H. Timmermans. A Model of Pedestrian Route Choice and Demand for Retail Facilities Within Inner-City Shopping Areas. *Geographical Analysis*, Vol. 18, 1986, pp. 115-128.
16. Willis, P., and D. Hodge. Trip Chaining in the Puget Sound Region: New Perspectives on Travel Behavior. Presented at 69th Annual Meeting of the Transportation Research Board, Washington, D.C., January 1990.
17. *Chicago Area Transportation Study, 1970 Household Travel Surveys*. Final Report. Creighton Hamburg Incorporated, Chicago, Ill., July 6, 1971.
18. CATS Household Travel Surveys, Vol. 1 to 9. Documentation for each of the areas surveyed, Chicago Area Transportation Study, Vol. 1, Sept. 1989, through Vol. 9, Aug. 1993.
19. Kim, H., J. Li, S. Roodman, A. Sen, S. Sööt, and E. Christopher. Factoring Household Travel Surveys. Presented at 72nd Annual Meeting of the Transportation Research Board, Washington, D.C., January 1993.
20. Chicago Area Transportation Study. *Transportation Facts-- A Focus on Census Work Trips*, Vol. 10, No. 3, June 1993.
21. O'Kelly, M., and E. Miller. Characteristics of Multistop Multipurpose Travel: An Empirical Study of Trip Length. In *Transportation Research Record 976*, TRB, National Research Council, Washington, D.C., 1984, pp. 33-39.
22. Hu, P., and J. Young. *Summary, of Travel Trends, 1990 Nationwide Personal Transportation Survey*. FHWA, U.S. Department of Transportation, March 1992, p. 43.

Publication of this paper sponsored by Committee on Passenger Travel Demand Forecasting.

Estimation of Travel Demand Models with Grouped and Missing Income Data

CHANDRA BHAT¹

A method to impute a continuous value for household income from grouped and missing income data for use as an explanatory variable in travel demand estimation was developed. Many data sets collect income in a discrete number of categories or in grouped form to simplify the respondent's task and to encourage a response. In spite of such grouped data collection, many respondents refuse to provide information on income, leading to missing income values. The issue of constructing a continuous measure of income from grouped and missing income data that, when used in travel demand models as an explanatory variable, enables consistent estimation of the model parameters is addressed.

Household income is an important sociodemographic explanatory variable in travel demand models such as car ownership models (1), trip generation models (2), and mode choice models (3). In almost all transportation data sets and in many other data sets (4) household income, an inherently continuous variable, is measured in a discrete number of categories or intervals; that is, it is measured in grouped form (e.g., between \$15,000 and \$30,000). The income question is also notorious for its high nonresponse rates, leading to missing income observations in most data sets.

Income is measured in grouped form for two related reasons. First, such a measuring scale provides a greater degree of protection of confidentiality compared with a continuous measure (the degree of protection being a function of the size of income intervals), thereby increasing response rates (5). Second, it renders the sensitive income question relatively innocuous during survey administration. Questions that seek a continuous measure on income can offend respondents, particularly in a telephone survey or in a personal interview survey in which respondents are put "on the spot."

Although income is measured in grouped form, it is the continuous measure of income (or some function of this continuous measure) that frequently appears as an explanatory variable in travel demand models. It is important that this continuous measure be a reliable measure of the true income value to enable the development of an accurate and reliable relationship between travel demand variables and their explanatory variables and thus facilitate good prediction of travel demand variables [the research by Hamburg et al. (6) indicates that the estimates in a travel demand model are highly sensitive to the accuracies of sociodemographic input variables and emphasizes the need for accurate measures of the input variables]. This paper proposes a method for constructing such a continuous measure of income for all observations in a cross-sectional data set with grouped and missing income data.

¹ Department of Civil Engineering and Environmental Engineering, University of Massachusetts, Amherst, Mass. 01003.

The next section of this paper discusses the motivation for developing methods to explicitly accommodate the grouped and missing nature of income data in travel demand modeling. The subsequent section presents the need to develop a model relating household income and factors affecting household income to impute a continuous income measure from grouped and missing income data for use as an exogenous variable in travel demand models. The following section advances an econometric framework used to impute a continuous income measure through the development of a model relating income to variables influencing income. Empirical results obtained by using a Dutch data set are then presented. The final section provides a summary of the research and highlights important findings.

MOTIVATION FOR TREATMENT OF GROUPED AND MISSING INCOME DATA

The motivation for the treatment of grouped and missing income data originates from the need to develop a consistent relationship between travel demand variables and their explanatory variables (including income). The dependent variable in the demand model may be an observed continuous variable such as trip generation or a latent continuous variable that is a reflection of an observed discrete choice decision such as utilities in the case of a mode choice decision or car ownership propensity in the case of an ordered car ownership model. Unfortunately current procedures for constructing a continuous measure from grouped data and commonly used techniques for handling missing income data do not enable consistent estimation of travel demand models. This inconsistency in commonly used demand estimation procedures without and with missing income data is discussed below.

Commonly Used Estimation Procedures

Grouped Without Missing Income Data

Commonly used estimation procedures construct a continuous value of income from grouped data by assigning the midpoint of each of the income threshold bounds that determine each category to each observation in that category. If the threshold bounds for income category j are a_{j-1} (the lower bound) and a_j (the upper bound), then a continuous income value ζ is constructed for all observations in category j as

$$\zeta_i | (\text{income category} = j) = \frac{a_{j-1} + a_j}{2} \quad (1)$$

In the case of the two categories at either end of the income spectrum, an arbitrary truncation point is used as the representative value.

This midpoint method of constructing a continuous measure from grouped data has serious limitations. Consider an underlying linear regression between a demand variable y , and the actual (but unobservable) income variable I_i^* as follows [the following presentation is based on Hsiao (7) and is confined to the case when the dependent demand variable is an observed continuous variable for ease in presentation]:

$$y_i = \alpha + \beta I_i^* + u_i \quad (2)$$

where

- i = index for observations,
- α and β = parameters to be estimated, and
- u_i = an error term.

Assume the standard regression conditions that u_i is an independent and identically distributed (iid) random error term with a mean of zero and I_i^* is uncorrelated with the error term. If the actual income value I_i^* for an observation is replaced by the midpoint of the corresponding income category, the regression may be rewritten as

$$y_i = \alpha + \beta \zeta_i + v_i \quad (3)$$

where $v_i = u_i + \beta(I_i^* - \zeta_i)$. In this case when the midpoint values are used, the coefficient of β is given by (using Equation 2)

$$\hat{\beta}_{mid} = \frac{\sum_i (y_i - \bar{y})(\zeta_i - \bar{\zeta})}{\sum_i (\zeta_i - \bar{\zeta})^2} = \beta \frac{\sum_i (I_i^* - \bar{I}^*)(\zeta_i - \bar{\zeta})}{\sum_i (\zeta_i - \bar{\zeta})^2} \quad (4)$$

To simplify this expression write the actual (but unobserved) continuous income I_i^* for an observation i falling in the grouped income category j as the sum of three components: the midpoint of the category j , ζ_j , as computed in Equation 1; an error term τ_i representing the difference between the expected value of I_i^* given that it falls in category j (or the expected value of the marginal distribution of the continuous income variable between the threshold bounds of category j) and the midpoint of category j ; and a random error term, w_i , representing the difference between the actual continuous income I_i^* and the expected value of I_i^* given that it falls in category j . That is,

$$I_i^* = \zeta_j + \tau_i + w_i \quad (5)$$

where $\tau_i = E[I_i^* | cat. j] - \zeta_j$ and $w_i = I_i^* - E[I_i^* | cat. j]$. By using Equation 5 one can

$$P \lim_{N \rightarrow \infty} \hat{\beta}_{mid} = \beta + \beta \frac{Cov(\tau_i, \zeta_j)}{Var(\zeta_j)} \quad (6)$$

write $I_i^* - \bar{I}^* = (\zeta_i - \bar{\zeta}) + (\tau_i - \bar{\tau})$. By substituting this expression into Equation 4 one can rewrite the least-squares estimate of β by the midpoint method as

Thus the parameter estimate on income obtained by the midpoint method converges to the actual value of 0 in the travel demand model if and only if $Cov(\tau_i, \zeta_i)$ converges to zero. However this will generally not be the case. The magnitude and direction of $Cov(\tau_i, \zeta_i)$ depend on the shape and distribution of the actual (but unobserved) income variable. Earlier studies (8,9) have indicated that a log-normal form is theoretically and empirically appropriate for the income distribution. $Cov(\tau_i, \zeta_i)$ is, in general, not equal to zero for a log-normal distribution. No general result regarding the direction and magnitude of $Cov(\tau_i, \zeta_i)$ (and therefore the direction and magnitude of the bias of the midpoint method) can be established for the log-normal distribution. A more definitive result can be established if it is assumed that I_i^* in Equation 2 represents the logarithm transformation of actual income. In this case I_i^* is normally distributed (since actual income is log normally distributed). Assuming small tail distributions, τ_i decreases from a positive value for the lower income categories (the expected value of the normal distribution between the threshold bounds of category j is greater than the midpoint) to a negative value for the higher income categories (the expected value of the distribution between the threshold bounds of category j is lower than the midpoint) as indicated by Haitovsky (10). On the other hand the midpoint of income categories increases as one proceeds from lower to higher categories. Thus the covariance term, $Cov(\tau_i, \zeta_i)$, is negative and the midpoint estimate $\hat{\beta}_{mid}$ in Equation 6 underestimates β .

The midpoint method leads to inconsistent parameter estimates (a parameter estimate $\hat{\beta}$ is said to be a consistent estimator of the true β if, as the sample size gets infinitely large, the probability that $|\beta - \hat{\beta}|$ will be less than any arbitrary small positive number approaches 1) in the travel demand model because τ_i is not equal to zero. However if a consistent imputed estimate of income (that is, a consistent estimate of the expected value of I_i^* given that it falls in category j) is used instead of the midpoints, τ_i is zero and one obtains consistent parameters in the travel demand model (the reader will observe that as the number of income categories increases, or more appropriately as the size of the income interval within each income category decreases, τ_i becomes closer to zero in the midpoint method and the inconsistency resulting from use of the midpoint method is reduced).

The results regarding the inconsistency of the midpoint method are generalizable to the case of many explanatory variables in the travel demand model. Specifically use of the midpoint income estimate as an explanatory variable leads to inconsistent parameter estimates on all of the explanatory variables in the model, not just the income variable (7).

Grouped with Missing Income Data

The discussion above assumed that there are no missing income observations. Now consider the limitations of commonly used methods when missing income data are present. Current methods adopt one of two strategies to estimate travel demand models from grouped and missing income data. The first strategy is to assign the midpoint of income categories for observations with observed (grouped) income values and to assign the average value of the midpoint estimates of the observed income observations to the missing income observations. As discussed earlier, the midpoint method does not provide consistent estimates of the travel demand model. Also this assignment of the average of observed income observations to missing income observations assumes that the average income of respondent households (i.e., households that report income) is identical to that of non respondent households (i.e., households that do not report income). This may not be true because of systematic variations in observed and unobserved characteristics affecting income earnings between members of respondent and nonrespondent households (11). Observed characteristics may include the education levels of the members of the household, whereas unobserved characteristics may include sensitivity to privacy and fear of governmental or other uses of the data. If such systematic variations are present between members of respondent and nonrespondent households, assigning the average income of respondent households to nonrespondent households is inappropriate and will further contribute to inconsistency in the parameter estimates of the demand model.

The second strategy for estimating travel demand models from grouped and missing income data is to assign the midpoint of income category thresholds for the observed (grouped) income data and to drop all missing income observations. It was already shown that the midpoint method provides inconsistent travel demand parameters. In addition another dimension of inconsistency arises when all missing income observations are dropped. If systematic variations in income level are present between respondent and nonrespondent households, then the relationship between independent variables and the travel demand variable for nonrespondents may be different from that for respondents. Thus the travel demand relationship obtained by dropping all nonrespondent households will not be a representative relationship for the entire population. This second strategy of dropping missing income observations also results in a loss of observations, resulting in inefficient estimation.

It is clear from the discussion above that commonly used procedures for dealing with grouped and missing income data are inadequate or waste valuable data. The next section discusses the need to develop a dependent income model, that is, a relationship between household income (the dependent variable) and a set of variables affecting household income (the independent variables), to impute a continuous income measure from grouped and missing data for use as an explanatory variable in travel demand models.

NEED FOR DISAGGREGATE INCOME MODEL FOR IMPUTING INCOME

This section discusses the need to develop a dependent income model to impute a continuous income measure. Cases in which there are no missing income data and in which there are missing income data are discussed.

No Missing Income Data

Earlier it was indicated that use of a consistent imputed estimate of income (that is, assigning to each observation falling in income category j the expected value of the income distribution bounded by the category thresholds) in a travel demand model provides consistent parameter estimates. This method assigns a single value to all income observations in a category. It does not use information on observed variables likely to affect income earnings (such as education level and number of employed adults in a household) that can help to differentiate among the incomes of different households within a particular grouped category. Developing a dependent income model (using the grouped observation on income) and combining the instrumental variable estimate of income from such a model with the information on income categories will enable construction of a consistent and efficient imputed income measure for use in travel demand models. The structure and estimation procedure for imputing income values from grouped data are discussed later in this paper.

Presence of Missing Income Data

The need to develop a dependent income model is critical when missing income data are present, since such a model is the only means of imputing an income measure for the missing data while at the same time accounting for any systematic variations in the observed characteristics (such as education level and number of employed adults) between respondent and nonrespondent households. The model should also account for systematic variations in unobserved characteristics between respondent and nonrespondent households. A consistent and efficient imputed estimate of income for use in travel demand models can be obtained from grouped and missing income data by combining the instrumental variable estimate of income from the model with information on whether a household responded to the income question or not and the income category in which a household's income falls (if the household responded). The structure and estimation procedure for imputing income values from grouped and missing income data are discussed later in this paper.

The discussion above emphasizes the need to develop a dependent income model to impute a continuous income estimate from grouped or grouped and missing income data for use as an exogenous variable in travel demand models. The remainder of this paper presents the econometric framework for imputing income through the development of a dependent income model and presents empirical results of the dependent income model and associated imputed estimates by using a Dutch data set.

ESTIMATION METHODOLOGY

The methodology used to develop a dependent income model and to impute a continuous income value from grouped and missing data in two stages is discussed in this section. In the first stage it is assumed that there are no missing income values. The methodology is then extended to accommodate missing income values in the second stage. The program routines for all estimations in this paper were written and coded by using the GAUSS matrix programming language.

No Missing Income Data

Assume that the actual but unobserved logarithm of household income, I_i^* , is a function of a vector X_i of exogenous variables as follows:

$$I_i^* = \gamma' X_i + \epsilon_i \quad (7)$$

where

- γ = vector of parameters to be estimated,
- X_i = vector of explanatory variables, and
- ϵ_i = a random disturbance term assumed to be homoscedastic, independent, and normally distributed with mean of zero and a variance of σ^2 (a logarithm form is adopted for the dependent income variable because as indicated earlier a log-normal form has been found to be theoretically and empirically appropriate for the income distribution).

The observed data on income indicate that they fall into a prespecific interval. The relationship between the grouped observed income data I_i and the continuous unobserved (log) income value I_i^* is written as follows:

$$I_i = j \quad \text{if } a_{j-1} < I_i^* \leq a_j, \quad j = 1, \dots, J, i = 1, \dots, N \quad (8)$$

where the a_j 's represent known threshold values (which represent the logarithm of the actual income threshold bounds) for each income category j . Representing the cumulative standard normal by Φ , the probability that household income falls in category j may be written from Equations 7 and 8 as

$$\text{Prob}(I_i = j) = \Phi\left(\frac{a_j - \gamma' X_i}{\sigma}\right) - \Phi\left(\frac{a_{j-1} - \gamma' X_i}{\sigma}\right) \quad (9)$$

Defining a set of dummy variables

$$M_{in} = \begin{cases} 1 & \text{if } I_i^* \text{ falls in the } j\text{th category} \\ 0 & \text{otherwise,} \end{cases} \quad (i = 1, 2, \dots, N, j = 1, 2, \dots, J) \quad (10)$$

the likelihood function for estimation of the parameters γ and σ is

$$\mathcal{L} = \prod_{i=1}^N \prod_{j=1}^J \left[\Phi\left(\frac{a_j - \gamma' X_i}{\sigma}\right) - \Phi\left(\frac{a_{j-1} - \gamma' X_i}{\sigma}\right) \right]^{M_{ij}} \quad (11)$$

Initial parameter values for the maximum likelihood search are obtained by assigning to each income observation its conditional expectation on the basis of the marginal distribution of I^* and regressing these conditional expectations on the vector of exogenous variables. The reader will note that the likelihood function of Equation 11 differs from that of the standard ordered probit model. In particular σ is unidentifiable and the threshold values (the a_j 's) are unknown parameters to be estimated in the ordered probit model. In contrast in the current model the threshold values are known, and (as a consequence) σ is identifiable.

Defining the standard normal density function by $\phi(\cdot)$, an imputed value for household (log) income may be computed for all the observations from the estimates of γ and σ obtained from maximizing the likelihood function in Equation 11. The imputed value for an income observation in category j may be computed by using the properties of doubly truncated univariate normal distributions (12) as follows:

$$\hat{I}_i^* | (X_i, I_i = j) = \hat{\gamma}' X_i + \hat{\sigma} \frac{\phi\left(\frac{a_{j-1,i} - \hat{\gamma}' X_i}{\hat{\sigma}}\right) - \phi\left(\frac{a_{j,i} - \hat{\gamma}' X_i}{\hat{\sigma}}\right)}{\Phi\left(\frac{a_{j,i} - \hat{\gamma}' X_i}{\hat{\sigma}}\right) - \Phi\left(\frac{a_{j-1,i} - \hat{\gamma}' X_i}{\hat{\sigma}}\right)} \quad (12)$$

These imputed values represent unbiased and consistent measures of (log) income and can be used as an explanatory variable in travel demand models (the imputed values are also guaranteed to fall within the lower and upper boundaries of the observed income categories). If an alternative function of income (other than the log function), $g(I_i^*)$, appears as the explanatory variable in the travel demand model, an imputed value may be computed as:

$$\hat{g}(I_i^*) = g(\hat{I}_i^*) \quad (13)$$

This imputed value of the function of (log) income is not unbiased, since in general the expected value of a continuous function of a variable is not equal to the function of the expected value of the variable. However it is consistent by Slutsky's theorem and thus will enable consistent estimation of travel demand models.

Presence of Missing Income Data

If missing income values are present in the data (as is almost always the case), one of two approaches may be used to construct a continuous value for all observations: (a) the naive approach or (b) the sample selection approach.

Naive Approach

The naive approach employs the method described above to estimate γ and σ by using only the observed (and grouped) income values. A continuous (log) income value is then imputed by using Equation 12 for observed income values and using $\hat{I}_i^* = \hat{\gamma}' X_i$ for missing income values. The naive approach accounts for systematic differences in the observed characteristics (represented by the X vector in Equation 7) that affect income between households that provide income and those that do not. However it fails to accommodate for systematic differences in the unobserved characteristics that affect income between respondent and nonrespondent households; that is, it ignores any "self-selection" in the choice of households to report income. Specifically unobserved factors that affect household income may also influence the decision of individuals (or households) to report income. For example it seems at least possible that households with above-average incomes, other things being equal, will be more reluctant than other households to provide information on income [Lillard et al. (11) indicate that this is so in their study of the 1980 Census Population Survey]. Because of this potential sample selection [see Mannering and Hensher (13) for a detailed review of sample selection-related issues], the naive approach will not, in general, provide consistent (continuous) estimates of income for observed or missing income data [the method proposed by Stem (14) for imputing income from grouped and missing income data falls under the naive approach]. To obtain consistent estimates the decision to report income should be considered endogenous, as discussed in the next section.

Sample Selection Approach

The sample selection approach uses two equations, one for income reporting and the other for household income, and accounts for the correlation in error terms between the two equations. Thus it accommodates systematic differences in unobserved characteristics between respondent and nonrespondent households. The model system is as follows:

$$r_i^* = \gamma_r' X_{ri} + \epsilon_{ri}, r_i = 1 \text{ if } r_i^* > 0 \text{ and } r_i = 0 \text{ if } r_i^* \leq 0 \quad (14)$$

$$\left. \begin{array}{l} I_i^* = \gamma_i' X_{li} + \epsilon_{li} \\ I_i = j, \text{ if } a_{j-1} < I_i^* \leq a_j \end{array} \right\} \text{observed only if } r_i^* > 0 \quad (15)$$

where

- r_i = observed binary variable indicating whether or not income is reported ($r_i = 1$ if income is reported and $r_i = 0$ otherwise),
 r_i^* = underlying continuous variable related to the observed binary variable r_i as shown above,
 X_{ri} and X_{li} = vectors of exogenous variables,
 γ_r and γ_l = vectors of parameters to be estimated, and
 ε_{ri} and ε_{li} = normal random error terms assumed to be independent and identically distributed across observations with a mean of zero and variance of one and σ_l^2 , respectively.

The error terms are assumed to follow a bivariate normal distribution (the author is not aware of any earlier application of sample selection in econometric literature in which the variable subjected to sample selection is observed only in grouped form).

The probability that income is observed and falls in income category j from the model system of Equations 14 and 15 is:

$$\text{Prob}(r_i = 1, I_i = j) = \Phi_2\left(\frac{a_j - \gamma_l' X_{li}}{\sigma_l}, \gamma_r' X_{ri}, -\rho\right) - \Phi_2\left(\frac{a_{j-1} - \gamma_l' X_{li}}{\sigma_l}, \gamma_r' X_{ri}, -\rho\right) \quad (16)$$

where ρ is the correlation between the error terms ε_{ri} and ε_{li} and Φ_2 is the cumulative standard bivariate normal function.

Defining a set of dummy variables M_{ij} as in Equation 10 for the observed income observations, the appropriate maximum likelihood function for estimation of the parameters in the model system is

$$\begin{aligned} \mathcal{L} = & \prod_{i=1}^N [1 - \Phi(\gamma_r' X_{ri})]^{1-r_i} \\ & \times \left\{ \prod_{j=1}^J \left[\Phi_2\left(\frac{a_j - \gamma_l' X_{li}}{\sigma_l}, \gamma_r' X_{ri}, -\rho\right) - \Phi_2\left(\frac{a_{j-1} - \gamma_l' X_{li}}{\sigma_l}, \gamma_r' X_{ri}, -\rho\right) \right]^{M_{ij}} \right\}^{r_i} \quad (17) \end{aligned}$$

Initial start values for the *ML* iterations are obtained by assigning to each reported income observation its conditional expectation on the basis of the marginal distribution of the underlying latent continuous variable I_i^* . These values are now treated as the actual continuous (log) income values, and a Heckman's two-step method (15) is applied for sample selection models to obtain start values for the parameters.

The continuous value of (log) income for households that reported income may be computed from the parameter estimates obtained from maximizing Equation 17. By using the properties of doubly truncated bivariate normal distributions (16) and defining the following quantities,

$$m = \frac{a_j - \hat{\gamma}'_l X_{li}}{\hat{\sigma}_l}$$

$$k = \frac{a_{j-1} - \hat{\gamma}'_l X_{li}}{\hat{\sigma}_l}$$

$$g = \frac{\hat{\gamma}'_r X_{ri} + k\hat{\rho}}{\sqrt{1 - \hat{\rho}^2}}$$

$$r = \frac{k + \hat{\gamma}'_r X_{ri} \hat{\rho}}{\sqrt{1 - \hat{\rho}^2}}$$

$$s = \frac{m + \hat{\gamma}'_r X_{ri} \hat{\rho}}{\sqrt{1 - \hat{\rho}^2}}$$

one can write

$$\begin{aligned} \hat{I}_i^*|(X_{ri}, X_{li}, r_i = 1, I_i = j) &= \hat{\gamma}'_r X_{li} \\ &+ \hat{\sigma}_i \frac{\phi(k)\Phi(g) - \phi(m)\Phi(h) + \hat{\rho}\phi(-\hat{\gamma}'_r X_{ri})[\Phi(s) - \Phi(r)]}{\Phi_2(\hat{\gamma}'_r X_{ri}, m, -\hat{\rho}) - \Phi_2(\hat{\gamma}'_r X_{ri}, k, -\hat{\rho})} \end{aligned} \quad (18)$$

The above expression collapses to Equation 12 if the correlation between the error terms in the reporting equation and the income equation is zero.

The continuous value of (log) income for households that did not report income may be imputed as follows:

$$\hat{I}_i^*|(X_{ri}, X_{li}, r_i = 0) = \hat{\gamma}'_l X_{li} - \hat{\rho}\hat{\sigma}_l \left(\frac{\phi(\hat{\gamma}'_r X_{ri})}{1 - \Phi(\hat{\gamma}'_r X_{ri})} \right) \quad (19)$$

EMPIRICAL RESULTS

This section discusses the data used to develop the dependent income model and to impute income from grouped and missing income observations and also presents estimation results.

Data

The data source used in the present study is from a Dutch National Mobility Survey. The survey involved weekly travel diaries and household and personal questionnaires collected during the spring of 1988 [for a detailed description of this survey see van Wissen and Meurs (17)]. The sample included 889 households, 55 of which have missing income data. Household income was available in three categories (for the observed income observations) in the data: (a) less than or equal to 24,000 guilders, (b) from 24,001 to 28,000 guilders, and (c) greater than 38,000 guilders.

Empirical Specification and Results

The variables considered in the income reporting equation and household income equation are listed in Table 1. They included household age and education (see definitions in Table 1), number of employed adults in the household, number of kids in the household, an indicator of whether the household has a "returning" young adult, and unemployment rate in the municipality of household residence. The household age variables enable nonlinear estimation of the age effect on income reporting and income earnings. The education variables indicate the effect of different levels of education of the adults in the household relative to that for households with one or more adults with primary education.

TABLE 1 Exogenous Variables in Model

TABLE 2 Estimation Results

Equation	Variable	The naive approach		The sample selection approach	
		Coefficient	t stat.	Coefficient	t stat.
Reporting equation	constant	-	-	2.218	1.77
	household age				
	entire range	-	-	0.002	0.04
	> 35 years	-	-	0.065	0.90
	> 45 years	-	-	-0.141	-2.31
	household education				
	secondary/high	-	-	-0.762	-3.50
	high	-	-	-1.114	-4.54
Income equation	number of kids	-	-	-0.150	-1.27
	RYA family	-	-	-0.759	-2.08
	constant	10.053	57.35	10.051	56.75
	household age				
	entire range($\times 10^{-1}$)	-0.006	0.12	-0.002	0.05
	> 35 years	0.012	1.54	0.011	1.42
	> 45 years	-0.006	-0.94	-0.004	-0.62
	household education				
	secondary	0.188	6.63	0.188	6.60
	secondary/high	0.364	10.67	0.382	11.05
	high	0.414	10.12	0.446	10.59
	number of employed adults	0.258	10.83	0.258	10.73
	unemployment rate	-1.125	-3.20	-1.117	-3.19
	σ	0.267	19.36	0.273	18.15
Correlation term	ρ	-	-	-0.694	-2.33
# of observations		834		889	
Log Likelihood (slopes = 0, $\rho = 0$)		-800		-1006	
Log Likelihood (convergence)		-621		-797	

The naive method and the sample selection method were used to estimate the parameters in the household income equation. The naive method estimates parameters from observed income observations b), using Equation 11, whereas the sample selection method estimates parameters from all observations by using Equation 17. The results are shown in Table 2. The naive method estimates only the income equation, whereas the sample selection method estimates both the reporting equation and the income equation and accounts for the correlation in unobserved factors that affects these equations simultaneously. In both models the level of household education and the number of employed adults have a positive effect on income. The magnitudes of the parameters on household education are consistent with the expectation that higher levels of education have a greater effect on income. The unemployment rate in the municipality of the household residence has a significant negative effect. The reporting equation estimation results in the sample selection model indicate that households with older adults, households whose individuals have a high level of education, and households with a returning young adult have a significant negative effect on reporting. Thus there are systematic differences in the observed characteristics between households that report income and those that do not.

The magnitude and significance of the correlation term ρ in the sample selection model indicate that there is a significant (and rather high) negative correlation in the unobserved factors that affect the reporting equation and the household income equation; that is, households that did not report their incomes were, all observed characteristics being equal, likely to have higher incomes than households that reported their incomes. This indicates that the naive method provides biased and inconsistent estimation results. In particular the naive method tends to underestimate the magnitudes of parameters on the exogenous variables that have a positive effect on income and tends to overestimate the magnitudes of parameters on the exogenous variables that have a negative effect on income in the income equation because of the negative correlation between the error terms in the reporting and the income equations (although the difference in coefficient estimates between the naive and the sample selection approaches appears to be small, the reader should note that the dependent variable is the logarithm of income, and thus even small coefficient differences could translate into moderate differences with respect to income earnings; the small coefficient differences may also be attributable to the small number of missing income observations in the current data set).

TABLE 3 Mean Values of Imputed Income

Category	Mean imputed (log) income		
	Midpoint Approach	Naive Approach	Sample Selection Approach
Households which reported income (respondent households)	10.419 (0.319)	10.550 (0.245)	10.528 (0.317)
Households which did not report income (non-respondent households)	10.419 (0.000)	10.713 (0.280)	11.036 (0.240)

Note: Numbers in parentheses are centered standard deviations.

The mean values of imputed (log) income for households that reported income and those that did not report income obtained by using the midpoint method, the naive method, and the sample selection method are shown in Table 3. The mean values for the midpoint method depend on the representative value used for the lowest and the highest income categories. In the computations shown in Table 3 a value of log (15,000) was assigned for the "less than or equal to 24,000 guilders" category and a value of log (43,000) was assigned for the "greater than 38,000 guilders" category. The inconsistency and the ad hoc nature of the midpoint method of imputing income were discussed above. Furthermore the mean value of imputed (log) income was identical for both respondent and nonrespondent households by the midpoint method because the midpoint method does not account for systematic -variations in the observed and unobserved characteristics that affect income between respondent and nonrespondent households.

The naive method accounts for systematic variations in observed characteristics between respondent and nonrespondent households. The higher mean estimate for nonrespondent households compared with that for respondent households indicates that nonrespondent households have higher values than respondent households for the observed characteristics that increase income. This is readily observed in the reporting equation estimates of the sample selection model in Table 2, which indicate that nonrespondent households are characterized by adult members with a higher education level than those of adult members in respondent households.

The sample selection method accounts for systematic variations in the observed and unobserved characteristics that affect income between respondent and nonrespondent households. The difference in the mean value of imputed (log) income for respondent and nonrespondent households between the sample selection and naive approaches comprises two components. The first component is an underestimation of income by the naive method on the basis of the observed characteristics that affect income because of the biases in parameter estimates of the naive approach in Table 2. This first component leads to an increase in imputed (log) incomes for both respondent and nonrespondent households in the sample selection method compared with those in the naive method. The second component is the effect of the unobserved characteristics that affect reporting status and income. It leads to a decrease in imputed (log) income for respondent households and an increase for nonrespondent households. The naive method does not consider this second component; only the sample selection model does. The difference in the mean value of imputed (log) income between the sample selection and naive approaches is small for respondent households because the two components mentioned above act in opposite directions and tend to offset each other. On the other hand the main value of imputed (log) income from the sample selection approach is substantially larger than that from the naive approach for nonrespondent households because the two components mentioned above reinforce each other. Aside from the magnitude of the difference between the estimates of the sample selection and the naive approaches, however, the naive approach provides inconsistent imputed estimates both for respondent and for nonrespondent households because the correlation in the unobserved factors that affect reporting status and income earnings is significantly different from zero in Table 2. In general the sample selection method is the only approach that provides consistent imputed income estimates from grouped and missing income data.

CONCLUSION

This paper developed a methodology for imputing a continuous value of income from grouped and missing income data for use as an explanatory variable in travel demand models. The method was applied to data from the Dutch National Mobility Survey. In addition to indicating the applicability of the procedure developed in the paper to accommodating grouped and missing data, the results show that there are systematic differences in observed and unobserved characteristics between households that report income and households that do not. Failure to accommodate for this sample selection results in biased and inconsistent imputations. Use of such inconsistent imputed income values as an explanatory variable will result in unreliable travel demand models.

The methodology developed in this paper is particularly relevant because almost all transportation-related data bases record income in grouped form and because there is a trend for an increasing percentage of respondents to refuse to provide income information in travel and travel-related surveys (11). The methodology developed in the paper is easy to apply and has been coded for use with the GAUSS programming language.

ACKNOWLEDGMENTS

The author would like to thank Frank Koppelman and three anonymous referees for useful suggestions on previous versions of this paper.

REFERENCES

1. Golob, T. F. The Dynamics of Household Travel Time Expenditures and Car Ownership Decisions. Presented at the International Conference on Dynamic Travel Behavior Analysis, Kyoto, Japan, July 1989.
2. Meurs, H. Dynamic Analysis of Trip Generation. Presented at the International Conference on Dynamic Travel Behavior Analysis, Kyoto, Japan, July 1989.
3. Beggan, J. G. *The Relationship Between Travel Activity, Behavior and Mode Choice for the Work Trip*. M.S. thesis. Transportation Center, Northwestern University, Evanston, Ill., 1988.
4. Stewart, M. B. On Least Squares Estimation When the Dependent Variable Is Grouped. *Review of Economic Studies*, 1983, pp. 737-753.
5. Churchill, G. A., Jr., *Marketing Research: Methodological Foundations*. The Dryden Press, Chicago, 1983.
6. Hamburg, J. R., E. J. Kaiser, and G. T. Lathrop. *NCHRP Report 266: Forecasting Inputs to Transportation Planning*. TRB, National Research Council, Washington, D.C., 1983.
7. Hsiao, C. Regression Analysis with a Categorized Explanatory Variable. In *Studies in Econometrics, Time Series, and Multivariate Statistics*. Academic Press, Incorporated, New York, 1983.

8. Aitchison, J., and J. A. C. Brown. *The Lognormal Distribution with Special Reference to Its Uses Economics*. Cambridge University Press, Cambridge, 1976.
9. Mincer, J. *Schooling, Experience and Earnings*. National Bureau of Economic Research, New York, 1974.
10. Haitovsky, Y. *Regression Estimation from Grouped Observations*. Hafner Press, New York, 1973.
11. Lillard, L., J. P. Smith, and R. Welch. What Do We Really Know About Wages? Importance of Nonreporting and Census Information. *Journal of Political Economy*, Vol. 94, No. 31, 1986, pp. 489-506.
12. Johnson, N., and S. Kotz. *Distributions in Statistics: Continuous Multivariate Distribution*. John Wiley & Sons, Incorporated, New York, 1972.
13. Mannering, F., and D. A. Hensher. Discrete/Continuous Econometric Models and their Applications to Transport Analysis. *Transport Reviews*, Vol. 7, No. 3, 1987, pp. 227-244.
14. Stern, S. Imputing a Continuous Income Variable from a Bracketed Income Variable with Special Attention to Missing Observations. *Economic Letters*, Vol. 37, 1991, pp. 287-291.
15. Heckman, J. J. Sample Selection Bias as a Specification Error. *Econometrica*, Vol. 47, 1979, pp. 153-161.
16. Shah, S. M., and N. T. Parikh. Moments of Singly and Doubly Truncated Standard Bivariate Normal Distribution, *Vidya*, Vol. 7, 1964, pp.
17. van Wissen, L., and H. J. Meurs. The Dutch National Mobility Panel: Experiences and Evaluation. *Transportation*, Vol. 16, No. 2, 1989.

Publication of this paper sponsored by Committee on Passenger Travel Demand Forecasting.

Improved Kalman Filtering Approach for Estimating Origin-Destination Matrices for Freeway Corridors

NANNE J. VAN DER ZIJPP AND RUDI HAMERSLAG¹

The estimation of origin-destination (OD) matrices for freeway corridors by using inner-link induction loop data is examined. A trip generation model is used, and various parameter optimization and statistics-based methods are examined to estimate the split parameters in the model. A Kalman-based method that uses the model-predicted link-flow variances and covariances while processing the measurements is described. A simple but effective solution to the problem of initializing the Kalman filter and imposing the natural constraints to the estimates is presented. The resulting method is tested on both simulated and observed data and is compared with other methods such as least squares and constrained optimization, showing that the Kalman-based method leads to the best results.

Vehicle movement estimates are generally summarized in origin-destination (OD) tables. These tables contain the number of trips for each combination of origin and destination. For a freeway system origins correspond with on-ramps (entrances), whereas destinations relate to off-ramps (exits). Dynamically updated OD tables are required for various strategies aimed at optimal usage of existing freeway capacity. Examples of such strategies are ramp metering, route guidance, and incident management. Often induction loop data are the only continuously updated source of information, producing the number of observed vehicles per time slice. Induction loops generate an abundance of traffic counts. To be able to analytically calculate an OD table within a time slice, however, additional techniques are necessary. A first example of such a technique is the use of a traffic model that defines explicit relationships between OD flows. A second example is the use of an a priori trip table. The distance to this a priori trip table, according to some criterion, is minimized by using traffic counts as a boundary condition. Examples of these approaches can be found in Cascetta and Nguyen (1), Hamerslag and Immers (2), Bell (3), Hendrickson and McNeil (4), and van Zuylen and Willumsen (5).

Although the use of such techniques when applied to aggregated data sets can be well defended, it is questionable whether the inherent assumptions of the above-mentioned techniques are valid when applied to subnetworks like intersections or freeway corridors. First, these subnetworks contain neither real origins nor real destinations. Second, because of low aggregation levels, stochastic influences are likely to be dominant.

Therefore in this paper a class of OD estimators that works with a weaker assumption, the assumption of constant split ratios, is studied. According to this assumption for each entrance the *fractions* of traffic destined for a certain exit can be assumed to be changing slowly or even remain constant. This assumption changes the underspecific problem into an overspecified problem.

¹ Faculty of Civil Engineering, Delft University of Technology, P.O. Box 5048, 2600 GA Delft, The Netherlands.

The split ratio approach was first introduced by Cremer and Keller (6), who used a recursive formula to estimate the unknown split proportions. Since then various techniques have been used to estimate the split proportions. First, the correlation procedure was proposed by Cremer (7). This procedure is equivalent to the least-squares method. Later the method was improved by Cremer and Keller (8), who used constrained optimization. Simultaneously Kalman filtering was applied to this problem by both Cremer and Keller (8) and Nihan and Davis (9). Finally maximum likelihood approaches have been employed by Nihan and Davis (10) and Bell et al. (11). A combination of split ratio and modeling approaches can be found in Keller and Ploss (12), whereas Bell (13) added the problem of platoon dispersion.

The problem statement used in this paper will show many similarities to the problem statements used in the above-mentioned work. Three new elements are added, however. First, the measurement vector contains not only exit volumes but can also contain inner-link volumes. Second, the split parameters are interpreted as split probabilities rather than fixed fractions of entering volumes in a trip generation model. The third addition is the incorporation of a time shift in the problem. Entrance volumes and measurements from all locations are processed simultaneously. Therefore each measurement must be processed with a delay for all measurements to refer to one set of split parameters.

The main problem with the processing of inner-link volumes is the strong measurement dependency due to redundancy. To adequately describe the properties of the system and its measurements the trip generation model presented by van der Zijpp and Hamerslag (14) is used. This model distinguishes between split probabilities and split proportions, an idea already used by Davis and Nihan (15) and Davis (16) for static OD estimation. The trip generation model describes not only how split probabilities change through time but also the choice of destination as a random choice process and which noise is involved when monitoring entering traffic and inner-link volumes.

For prediction purposes the split probabilities have more significance than the split proportions. The OD estimation problem is therefore converted into the estimation of the split *probabilities* in the trip generation model. For this purpose least squares, constrained optimization, maximum likelihood, and Kalman filtering have been considered. Each method is described in terms of the variables used in the problem statement, and when necessary computational aspects are discussed. From these methods only the Kalman filter approach and the maximum likelihood approach allow the specification of dependency between measurements. From these two the Kalman filter has been selected because it is the only method for which tractable expressions for the dependent measurement case could be derived.

Several problems, however, hinder the straightforward application of Kalman filter to the problem of estimating the split probabilities from induction loop data. First, since the Kalman filter is a recursive method, a set of initial conditions needs to be available. Second, the measurement properties need to be defined, since noise occurs because of differences between split probabilities and split proportions and inaccuracies in induction loop observations. Finally there are several equality and inequality constraints that apply to the split probabilities. For each entrance split probabilities must not only add up to 1 but each individual split probabilities must also be nonnegative and less than 1. Depending on how these problems are solved one can expect a Kalman filter to do better or worse. The results presented by Cremer and Keller (8), for

example, suggest that constrained optimization gives better results than Kalman filtering, at the cost of high computation times.

The section Improved Kalman Filtering Approach describes a solution to each of these problems, resulting in an improved Kalman-based method. The method is tested against constrained optimization and least squares by using both simulated and empirical data. The test results are included in the last section.

PROBLEM STATEMENT

For the problems treated in this paper route choice is supposed to play no role, although this is not really a constraint of the methods under consideration; see for example Davis (16). All implemented methods take nonzero travel times into account. The problem of determining the delays is treated at the end of this section. For simplicity of notation the travel times are not mentioned in the equations.

Notation

The definitions of the terms used in the equations are as follows:

- $q(t)$ = vector of length m whose elements $q_i(t)$ are the observed volumes at entrance i that are processed during interval t .
- $y(t)$ = vector of length p whose elements $y_h(t)$ are the counted volumes at location h that are processed during interval t .
- $B(t)$ = $m \times n$ matrix whose elements $b_{ij}(t)$ are the proportion of trips leaving i destined for j . Let $b_{i:}(t)$ represent row i of $B(t)$, that is, the split parameters associated with entrance i . Then $b(t) = [b_{1:}'(t) \ b_{2:}'(t) \ \dots \ b_{m:}'(t)]'$ is defined as a vector of length $m \times n$ that contains the elements of $B(t)$ row by row.
- $F(t)$ = $m \times n$ matrix whose elements $f_{ij}(t)$ give the flow from i to j . Let $f_{i:}(t)$ define row i of $F(t)$. Then $f(t) = [f_{1:}'(t) \ f_{2:}'(t) \ \dots \ f_{m:}'(t)]'$ is a vector of length $m \times n$ that contains the elements of $F(t)$ row by row.

Trip Generation Model

The problem is to estimate the unknown parameters $B(t)$. Referring to van der Zijpp and Hamerslag (14), we argue that $b_{ij}(t)$ should be considered the *probability* that a vehicle will leave the network at exit j given the fact that it originated from entrance i . Such a probability does not really exist, but for our purposes drivers selecting randomly their destination upon entrance onto the network is an acceptable model of the system.

Working with split probabilities rather than split proportions has two advantages. First, the assumption that $b(t)$ is a slowly moving process can be better defended here since the randomly triggered difference between the split proportions and split probabilities is eliminated. Second,

$$b_{ij} \triangleq P[\text{exit at } j | \text{enter at } i] \quad (1)$$

some useful properties of the measurements can be derived, such as variances and covariances given a set of split probabilities. Hence from now on we use the following definition for the split parameters:

By definition the following constraints apply to the split parameters:

$$0 \leq b_{ij}(t) \leq 1 \quad i = 1 \dots m, j = 1 \dots n \quad (2)$$

$$\sum_{j=1}^n b_{ij}(t) = 1 \quad i = 1 \dots m \quad (3)$$

Like in Nihan and Davis (9) we refer to these constraints as the natural constraints. The split parameters are assumed to vary slowly over time, driven by a zero mean drift parameter $w(t)$:

$$b(t) = b(t-1) + w(t) \quad (4)$$

Another aspect we would like to consider is that all volumes are observed with noise because of inaccuracy of the induction loop observations. Introduce $q^*(t)$ and $y^*(t)$ as the vectors of real input and inner-link volumes, whereas $q(t)$ and $y(t)$ are the measured values. All noise components are considered to be independent and zero mean and have variances σ_q^2 or σ_y^2 . Therefore,

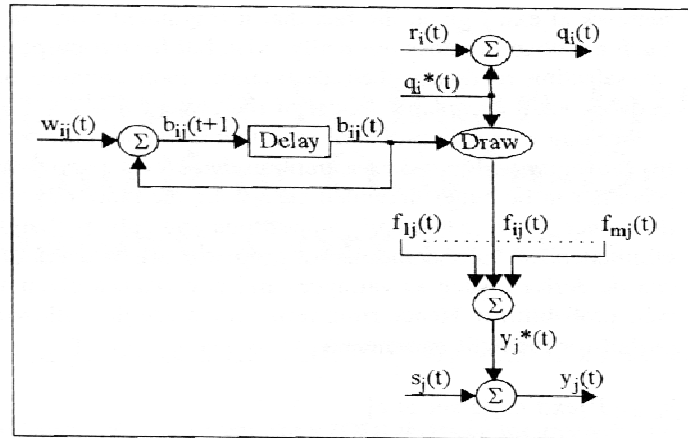
$$\begin{aligned} q(t) &= q^*(t) + r(t) \\ E[r(t)] &= 0, E[r(t)r(t)'] = \sigma_q^2 I \end{aligned} \quad (5)$$

and

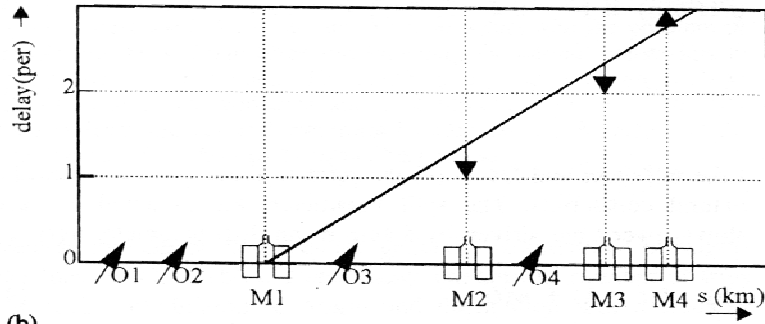
$$\begin{aligned} y(t) &= y^*(t) + s(t) \\ E[s(t)] &= 0, E[s(t)s(t)'] = \sigma_y^2 I \end{aligned} \quad (6)$$

Often the on-ramps are not monitored directly and one must calculate these entrance volumes by taking the difference of two consecutive inner-link volumes. Experiments have shown that neglecting noise in the entrance volume vector seriously affects the quality of the estimate, especially when the Kalman filter was applied. One reason for this is that the Kalman filter uses

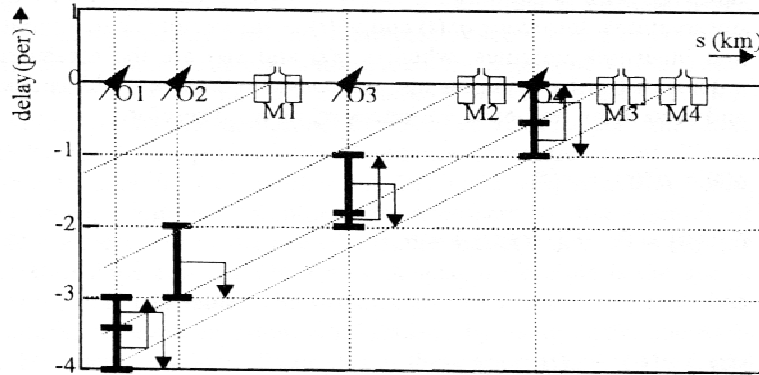
the entrance volume vector as a boundary condition. Therefore errors in the entrance volumes are subscribed to measurement noise. This causes a strong dependency among the elements of this noise vector.



(a)



(b)



(c)

FIGURE 1 Modeling assumptions: (a) trip generation model, (b) selecting simultaneously processed measurements, and (c) matching entrance volumes with measurements.

The above assumptions define the trip generation model that was presented earlier by van der Zijpp and Hamerslag (14). This model is summarized in Figure 1(a), which shows a system in which $b(t + 1)$ is obtained from $b(t)$ and drift variable $w(t)$, which will be considered a random input. The variables $b(t)$ are used as probabilities in a drawing process. For each entrance $q_i^*(t)$ experiments are done. The observed entrance volume vector $q(t)$ is obtained by taking the sum of $q^*(t)$ and $r(t)$, as described in Equation 5. The results of the drawing processes are merged into a link volume vector $y^*(t)$, to which (see Equation 6) a noise vector $s(t)$ must be added to obtain the measured values $y(t)$.

Measurements are available as traffic counts, which are obtained from induction loops. By definition these counts are linear combinations of flows. Since the route choice issue is neglected here, an OD flow is either totally or not at all contained in a measurement. Therefore,

$$y^*(t) = U'F(t) \quad (7)$$

with U denoting an $mn \times p$ matrix whose elements can either be one or zero, indicating that a flow does or does not contribute to the measurement. Note that this matrix does not depend on the time period. The transpose was used solely to keep conformity with literature. By using Equations 1 and 5 the following approximation for the flows can be derived:

$$f_{ij}(t) \approx q_i(t)b_{ij}(t) \quad (8)$$

Substituting this approximation in Equation 7 and combining this with Equation 6 allows us to calculate an $mn \times p$ matrix $H(t)$ with

$$y(t) = H'(t)b(t) + v(t) \quad (9)$$

This equation will later be referred to as the *measurement equation*. The vector $v(t)$ accounts for all measurement errors and the effects of the random selection process described in the trip generation model. The properties of this measurement error are discussed in the section Improved Kalman Filtering Approach.

Calculating Correct Time Delay

The measurements are processed with a time delay to let all measurements refer to the same set of split parameters and entrance volumes. However the time axis is divided into intervals, and the average travel times between entrances and measurement locations generally do not match the length of the intervals. To minimize errors a two-step process was followed.

The first step involves the selection of the measurements that will be processed simultaneously. To apply the natural constraint (Equation 3) to the estimate of $b(t)$, for each entrance this estimate must represent the splits during only one interval. To optimally fulfill this condition the relative travel times between the measurement locations are calculated and rounded to an integer number of intervals. This is illustrated in Figure 1(b), which shows the delay (in periods) as a function of the distance s (in kilometers). The locations of origins O1 through O4 and measurement locations M1 through M4 are indicated on the x-axis. The gradient of the line corresponds to the average speed. In the experiments described at the end of this paper this average could be derived directly from the input data, because measurements are carried out with double induction loops that monitor both intensity and speed.

The second step involves the selection of the corresponding entrance volumes. Since the average travel times do not exactly equal an integer number of periods, entrance volumes from at least two periods are assumed to contribute to a measurement. Therefore a weighted sum of the entrance volumes should be substituted in Equation 8. The weight factors can be determined from Figure 1(c). They correspond to the length of the vertical intervals in Figure 1(c). The arrows join the weight factors with the corresponding delay. By taking a weighted sum of two entrance volumes the optimal approximation of the entering volume during a certain period is obtained. However the entering traffic cannot be assumed to be evenly spread over time, and the travel times are not exactly known. Therefore it is inevitable that an error is introduced in the entering volume observation. For this reason the noise variable $r(t)$ was included in the trip generation model.

ESTIMATION OF SPLIT PARAMETERS

When we use the trip generation model described in the previous section, the problem of estimating the OD matrix is reduced to estimating the split parameters in the trip generation model. In this section five existing methods are described. Since these methods have been described in other contributions, this paper provides only a brief summary showing how the methods can be applied to problems in which the measurements contain inner-link volumes instead of exiting volumes only. The methods being considered here are the least-squares method, inequality-constrained least-squares method, constrained optimization, maximum likelihood, and Kalman filtering method. The first three methods can be classified as parameter optimization methods, whereas the other two are statistics-based methods.

Least-Squares Method

The least-squares method is aimed at solving the following problem:

$$\hat{b}(t) \sum_{k=1}^t ||y(k) - H'(k)\hat{b}(t)||^2 \quad (10)$$

By expanding this expression and setting the derivatives to $L(t)$ to zero, the least-squares estimate can be calculated by

$$\hat{b}(t) = \left[\sum_{k=1}^t H(k)H'(k) \right]^{-1} \left[\sum_{k=1}^t H(k)y(k) \right] \quad (11)$$

A unique solution is guaranteed if mn independent columns can be found in the matrices $H(l) \dots H(t)$. From Equation 11 it can be seen that it is possible to employ the least-squares method by using a constant amount of storage space by the following algorithm:

$$\begin{aligned} \hat{b}(t) &= HH_{tot}^{-1}(t)HY_{tot}(t) \\ HH_{tot}(t) &= HH_{tot}(t-1) + H(t)H'(t) \\ HY_{tot}(t) &= HY_{tot}(t-1) + H(t)y(t) \end{aligned} \quad (12)$$

By introducing a discounting factor the method can be adapted to track a time-varying $b(t)$. This transforms the problem into

$$\hat{b}(t) \sum_{k=1}^t \lambda^{t-k} \|y(k) - H'(k)\hat{b}(t)\|^2 \quad (13)$$

Again putting the derivatives to zero gives a minimum:

$$\hat{b}(t) = \left[\sum_{k=1}^t \lambda^{t-k} H(k)H'(k) \right]^{-1} \left[\sum_{k=1}^t \lambda^{t-k} H(k)y(k) \right] \quad (14)$$

This gives rise to the following algorithm:

$$\begin{aligned} \hat{b}(t) &= HH_{tot}^{-1}(t)HY_{tot}(t) \\ HH_{tot}(t) &= \lambda HH_{tot}(t-1) + H(t)H'(t) \\ HY_{tot}(t) &= \lambda HY_{tot}(t-1) + H(t)y(t) \end{aligned} \quad (15)$$

Inequality-Constrained Least-Squares Method

Equation 15 does not guarantee that the natural inequality constraint (Equation 2) is met. Imposing this condition would therefore improve the solution. On the other hand this would convert the problem from an unconstrained minimization into an inequality-constrained minimization problem:

$$\hat{b}(t) \sum_{k=1}^t \|y(k) - H'(k)\hat{b}(t)\|^2 \quad (16)$$

subject to

$$\hat{b}_l(t) \geq 0 \quad l = 1 \dots mn \quad (17)$$

This problem consumes much more computation time than the unconstrained problem. When solved by an interior steepest-descent method the computation times tend to be high because of the ill-conditioned matrix HH_{tot} . Not only does this hinder the testing of the method (an average test run would take 2 hr for the cases described in this paper) but also in case of real-time applications the duration of the computation could easily exceed the time available.

Therefore a less time-consuming algorithm is needed. For the time being the best results are obtained with an iterative algorithm that employs conjugate search directions that are projected on the feasible region when necessary. Moreover the searches are restricted to the feasible region, and the search direction is reset to steepest descent after each truncated search or change of active constraints. Calculation times are approximately 10 times longer than those by the straightforward matrix inversion method that could be used for the nonconstrained case. This suffices for problems of the size studied in this paper.

Constrained Optimization

If both the inequality and equality constraints (Equations 2 and 3) are imposed, an even better solution should be obtained. The satisfaction of the equality constraint (Equation 3) can be guaranteed by substituting the following in Equation 10:

$$\hat{b}(t) = b^0 + Gb^1(t) \quad (18)$$

with b^0 satisfying the equality constraints in Equation 3 and G being a $mn \times m(n-1)$ matrix chosen in such a way that $Gb^1(t)$ does not disturb the satisfaction of the equality constraints for all $b^1(t)$. Although many combinations of b^0 and G satisfy the necessary conditions, for practical reasons we use

$$b^0 = \begin{bmatrix} 0 \\ \dots \\ 0 \\ 1 \\ \dots \\ \dots \\ \dots \\ 0 \\ \dots \\ 0 \\ 1 \end{bmatrix}, G = \begin{bmatrix} & & & & & & & & & & \\ & & & & & & & & & & \\ & & I & & & & & & & & \\ & -1/n & -1/n \dots & -1/n & & & & & & & \\ & & & & \dots & & & & & & \\ & & & & & \dots & & & & & \\ & & & & & & \dots & & & & \\ & & & & & & & \dots & & & \\ & & & & & & & & I & & \\ & & & & & & & -1/n & -1/n \dots & -1/n & \end{bmatrix} \quad (19)$$

This substitution transforms the problem in Equation 10 into

$$\hat{b}(t) \sum_{k=1}^t ||y(k) - H'(k)(b^0 + Gb^1(t))||^2 \quad (20)$$

subject to

$$\begin{aligned} b_l^1(t) &\geq 0 \quad l = 1 \dots m(n-1) \\ \sum_{j=1}^{n-1} b_{(i-1)(n-1)+j}^1(t) &\leq 1 \quad i = 1 \dots m \end{aligned} \quad (21)$$

Solving this inequality-constrained problem and substituting the resulting $b^l(t)$ in Equation 18 gives a solution for $\hat{b}(t)$ that satisfies all required conditions. For solving the problem in Equation 20 the algorithms of the inequality-constrained problem can be used, although the projection of the search direction on the feasible region requires more computation time because of the nonorthogonal inequality constraints.

Maximum Likelihood

The previous methods can all be considered parameter optimization methods. They are designed to minimize the distance between measured and predicted values. Apart from these methods we

distinguish the statistics-based methods. These methods are defined in terms of the probability distributions related to the unknown parameters $b(t)$. The most common statistics-based method is the maximum likelihood (ML) technique. When applied to the problem of determining the split parameters in the trip generation model the ML solution would be defined by

$$\hat{b}(t) = \arg \max_{b(t)} P[y(1) \dots y(t) | b(t)] \quad (22)$$

Calculation of the ML solution normally requires the derivation of a probability distribution from the system shown in Figure 1(a), which is not tractable. Nihan and Davis (10) presented an ML approach that did not require this derivation by using the EM algorithm proposed by Dempster et al. (17). However this was done for the simplified system in which $b(t)$ was constant rather than slowly varying and in which no noise on the entrance volume observations was present. Moreover the resulting algorithm was nonrecursive.

Another ML approach has been presented by Bell et al. (11). This approach is fully disaggregate but is computationally too demanding to be useful in practice. So although ML estimators have desirable properties no ML estimator that suits our needs is available.

Kalman Filtering

Another statistics-based estimation technique is the Kalman filter. The Kalman filter is a widely applied method for parameter estimation in dynamic systems. Before a Kalman filter can be used two equations must be supplied: the state equation and the measurement equation. The state equation describes how the unknown parameters evolve through time. The measurement equation describes the relation between the unknown parameters and the measurement. In both equations it is possible to specify uncertainty by way of noise terms. In our case the state parameters represent the split probabilities. These parameters are assumed to change only slowly over time. Therefore we use the following state equation:

$$b(t) = b(t-1) + w(t) \quad (23)$$

The measurement equation describes the way the state parameters are observed. In this case we can use Equation 9:

$$y(t) = H'(t)b(t) + v(t) \quad (24)$$

In these equations $w(t)$ and $v(t)$ must be zero mean noise processes with known covariance matrices;

$$\begin{aligned} E[w(t)w(p)'] &= Q_t \delta_{tp} \\ E[v(t)v(p)'] &= R_t \delta_{tp} \end{aligned} \quad (25)$$

with δ_{tp} equal to 1 if t equals p and zero otherwise. On the basis of these equations and a knowledge of the covariance matrices a widely used estimation technique has been derived: the Kalman filter. A description of this technique can be found in many textbooks [see for example Anderson and Moore (18) and Catlin (19)]. The Kalman filter equations for the problems in Equation 23, 24, and 25 are:

$$\begin{aligned}\hat{b}(t) &= \hat{b}(t-1) + K_t[y(t) - H'(t)\hat{b}(t-1)] \\ K_t &= \sum_{t-1} H(t)[H(t)' \sum_{t-1} H(t) + R_t]^{-1} \\ \sum_t &= \sum_{t-1} - \sum_{t-1} H(t)[H(t)' \sum_{t-1} H(t) + R_t]^{-1} H'(t) \sum_{t-1} + Q_t\end{aligned}\quad (26)$$

These equations define a recursion that should be started with an initial estimate $\hat{b}(0)$ and an initial covariance matrix Σ_0 . Given the assumptions in Equations 23 through 25, the Kalman filter leads to the minimum variance linear estimator; that is, the estimate is a linear function of the measurements $y(1) \dots y(t)$, and the filter implicitly finds the matrix A and vector c that solve the following problem:

$$\min_{A, c} E[\|b(t) - A[y(1) \dots y(t)] - c\|^2] \quad (27)$$

Moreover this estimate can be shown to be unbiased. If besides earlier assumptions the noise terms and the initial state have Gaussian distributions the Kalman filter can be shown to produce unbiased estimates that have minimum variance over all estimators [see Anderson and Moore (18)].

The advantages of the Kalman filtering method are the computational efficiency of the method and the possibility of processing interdependent measurements. Other advantages are that the calculations can be done recursively and that together with the estimate for the split matrix a variance-covariance matrix is calculated. This matrix gives an indication of the reliability of the estimate.

IMPROVED KALMAN FILTERING APPROACH

Despite the nice theoretical properties of the method several problems hinder the straightforward application of a Kalman filter. The first problem is that no initial values $\hat{b}(0)$ and Σ_0 are available. Some experimenting shows that the problem of initializing the filter cannot be seen apart from a second problem: how to impose the natural inequality constraints in Equation 2. It seems natural to specify very large diagonal values of Σ_0 since this expresses a lack of information about $b(0)$ and results in forgetting the initial value $\hat{b}(0)$ as quickly as possible. On the other hand the initial variance is bounded above since the split parameters are bounded to the

interval [0,1]. Also specifying large initial variances results in many violations of the inequality constraints during the startup phase of the filter. The problem of dealing with these inequality constraints has already been treated by Nihan and Davis (9), who proposed several constraining algorithms. This paper shows that a much simpler and effective way of dealing with both initial conditions and inequality constraints is possible.

Another problem is the lack of information about the noise covariance matrices Q_i and R_i in Equation 25. The results produced by the Kalman filter strongly depend on these matrices. Therefore a good approximation of these matrices should improve the estimate. In this section the measurement noise covariance matrix is derived from the trip generation model shown in Figure 1(a). This derivation produces the matrix R_i as a function of the split probability $b(t)$. This is an approximation since only an estimate of $b(t)$ is available. The last problem treated in this section is the use of the natural equality constraints (Equation 3). In Nihan and Davis (9) a normalization procedure is used to impose these constraints. In this paper the natural equality constraint is imposed via the perfect measurement concept [see Anderson and Moore (18)]. The consequences for the method are discussed.

Initial Conditions and Inequality Constraints

The Kalman filter described in the previous section has one commonly recognized interpretation, that is, that of a linear minimum variance estimator. However the Kalman filter can also be interpreted as an example of Bayesian estimation [see Catlin (19)]. As shown by Maher (20), assuming a Gaussian a priori distribution of the state vector and performing a Bayesian update with a measurement that has a Gaussian distribution (conditionally to the state vector) leads to a Gaussian a posteriori distribution. The equations derived for the scalar measurement case in Maher (20) can be shown to match the Kalman filter measurement update equations. In van der Zijpp and Hamerslag (21) the results are generalized to nonconstant state parameters and nonscalar measurements. A central role in this derivation is played by Bayes rule:

$$p[b(t)|y(1)...y(t)] = \frac{p[y(t)|b(t), y(1)...y(t-1)]p[b(t)|y(1)...y(t-1)]}{p[y(t)|y(1)...y(t-1)]} \quad (28)$$

The validity of Bayes rule follows from the definition of conditional probability. According to Bayes rule the a posteriori distribution can be derived from the a priori distribution and the likelihood function of the measurement vector. Figure 2(a) illustrates the principle of Bayesian updating for a scalar Gaussian random variable and a scalar measurement. The a posteriori distribution is obtained by multiplying the a priori density and the likelihood function and normalizing the result.

Inequality Constraints

Since natural inequality constraints bound the split probabilities $b(t)$ to an mn -dimensional hypercube [0,1], the a priori probability function should be zero outside this hypercube. One way in which this can be achieved is by multiplying the a priori probability function with an indicator

function $I[0,1]()$. This function equals 1 if all elements of $b(t)$ satisfy the inequality constraints and is zero elsewhere. This leaves the shape of the distribution of $b(t)$ intact on the hypercube $[0,1]$, whereas it defines a zero probability elsewhere. To ensure that the new function integrates to unity the a priori function should be multiplied by a factor $F()$. $F()$ can be expressed as a function of $\hat{b}(t-1)$ and Σ_{t-1} . In this way we get a truncated MVN distribution.

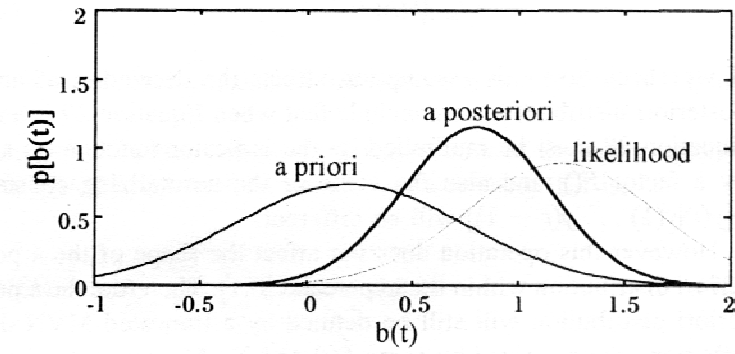
$$p[b(t)|y(1)\dots y(t-1)] = \frac{F(\hat{b}(t-1), \Sigma_{t-1})}{\sqrt{|\Sigma_{t-1} + Q_t|} (\sqrt{2\pi})^{mn}} \times \exp - \frac{1}{2} [b(t) - \hat{b}(t-1)]' (\Sigma_{t-1} + Q_t)^{-1} [b(t) - \hat{b}(t-1)] \times I_{[0,1]}[b(t)] \quad (29)$$

If we check how this assumption affects the derivation of an a posteriori distribution we conclude that when Equation 29 is used Equation 28 must be multiplied by the indicator function $I()$ and by a factor $F()$, and also the value of the normalizing constant $p[y(t)/(I(). \dots y(t-1))]$ will be different.

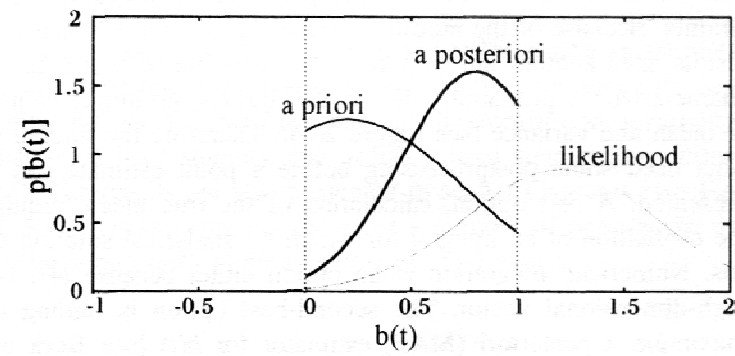
However this operation does not affect the shape of the a posteriori distribution within the hypercube $[0,1]$. Therefore the a posteriori distribution will still be defined by a truncated MVN distribution, characterized by some $\hat{b}(t)$ and Σ_t . Moreover the recursion that determines $\hat{b}(t)$ and Σ_t from $\hat{b}(t-1)$ and Σ_{t-1} has not been changed. Therefore the Kalman filter equations can be used without modification, despite the presence of inequality constraints. Because of the modified circumstances, the Kalman filter results need another interpretation. The variables $\hat{b}(t)$ and Σ_t still characterize the probability distribution but can no longer be used as mean and variance [see Figure 2(b)]. Therefore the filtered results need some postprocessing before a point estimate can be presented. A first option, calculation of the true mean, requires the evaluation of an integral for which no analytical solution exists. Numerical integration is no option either because $b(t)$ is a high-dimensional vector. The second-best option is finding the maximum a posteriori (MAP) estimator for $b(t)$ [see Beck and Arnold (22)]. This can be found by maximizing the a posteriori density of $b(t)$:

$$\min_{b(t)} (b(t) - \hat{b}(t))' \Sigma_t^{-1} [b(t) - \hat{b}(t)], 0 \leq b_i(t) \leq 1, i = 1, 2 \dots mn \quad (30)$$

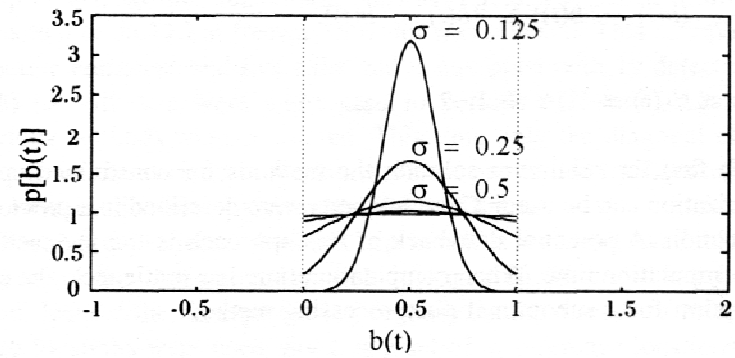
To find the minimum solution the methods for constrained optimization can be used. These methods were described in a previous section. A potential drawback of this approach is the increase in computation time. When computation time is a bottleneck one can option for a suboptimal postprocessing method.



(a)



(b)



(c)

FIGURE 2 Bayesian updating: (a) Bayesian update, (b) Bayesian update, truncated a priori distribution, and (c) uniform distribution approached by truncated normal distribution.

Initialization of Filter

In the foregoing we used the principle of Bayesian updating to derive a version of the Kalman filter that incorporates inequality constraints. As it will turn out, simultaneously we find a solution to the problem of initializing the Kalman filter. A common way of initializing a Bayesian filter when no a priori information is available is to use a uniform distribution. This expresses that, on the basis of the a priori information, every solution is equally likely. Working with the indicator function enables us to define an initial distribution that is arbitrarily close to the uniform distribution simply by defining Σ_0 as a diagonal matrix with very large diagonal elements. Figure 2(c) illustrates how a truncated Gaussian distribution approaches a uniform distribution if the variance increases.

Derivation of Measurement Noise Properties

The estimate obtained from a Kalman filter strongly depends on the assumed variance-covariance matrix for the measurement noise. Therefore in van der Zijpp and Hamerslag (14) such a matrix was derived on the basis of the trip generation model shown in Figure 1(a), which shows a system that is clearly different from the one described by the measurement Equation 24, since the measurements are numbers of successful experiments rather than linear combinations of the unknown parameters. However in terms of the expected value and the variance there is no difference between both systems. Therefore as far as the Kalman filter is concerned, we can treat the measurements from Figure 1(a) as if they were obtained from a linear system, as long as a covariance matrix R_t for the noise vector $v(t)$ is supplied.

A starting point for the derivation of such a matrix is the conditional distribution of the flows, given the entrance volume, $q_i^*(t)$, which is defined by a multinomial distribution:

$$P[f_{i1}(t) \dots f_{in}(t) | q_i^*(t)] = \frac{q_i^*(t)!}{\prod_{j=1}^n f_{ij}(t)!} \prod_{j=1}^n b_{ij}(t)^{f_{ij}(t)} \quad (31)$$

By combining this with Equation 5 it can be shown that the following equations define the covariance matrix for the measurement $y(t) = U'f(t)$ as a function of the split vector $b(t)$.

$$R_t = \text{cov}[y(t), y(t)] = U' \text{cov}[f(t), f(t)] U \quad (32)$$

with

$$\text{cov}[f_{ij}(t), f_{hk}(t)] = q_i(t) b_{ij}(t) \delta_{ih} \delta_{jk} + [\sigma_q^2 - q_i(t)] b_{ij}(t) b_{hk}(t) \delta_{ih} \quad (33)$$

Since the exact value of the split vector is unknown, the estimate of the split vector is used instead. The covariance matrix is therefore only an approximation to the true matrix.

Equality Constraints

Another way of improving the Kalman filter estimate is by imposing the natural equality constraints (Equation 3). For the purpose of imposing the natural equality constraints Niham and Davis (10) proposed a normalization procedure. Since that procedure was meant to act separately from the active parameter estimation method, it does not take full advantage of the possibilities of Kalman filtering.

Because the natural equality constraints are just another linear combination of the unknown split parameters, these constraints *mn* be imposed as measurements to the Kalman filter. These kinds of measurements are referred to as *perfect observations*, because no noise on these observations is present. In matrix notation,

$$e = F'b(t), e = \begin{bmatrix} 1 \\ 1 \\ \dots \\ 1 \end{bmatrix}, F' = \begin{bmatrix} 1 \dots 1 & & & \\ & 1 \dots 1 & & \\ & & \dots & \\ & & & 1 \dots 1 \end{bmatrix} \quad (34)$$

Anderson and Moore (18) show two ways to deal with these kinds Of observations. The first way is to reduce the order of the filter by an order *m* (*m* denotes the number of entrances). This can be done by a change of coordinate basis, similar to the one used while calculating the solution to the constrained optimization problem. The second way is to proceed as with any measurement by using a zero matrix for the measurement noise matrix. In this case a recursion similar to Equation 26 is valid. For ease of implementation the latter method was used in the study described in this paper.

Define $\hat{b}^+(t)$ and Σ_t^+ as the updated estimate and variance-covariance matrix after performing a measurement update by Equation 34. Now $\hat{b}^+(t)$ and Σ_t^+ are obtained via:

$$\begin{aligned} \hat{b}^+(t) &= \hat{b}(t) + K_t^+[e - F'\hat{b}(t)] \\ K_t^+ &= \Sigma_t F [F' \Sigma_t F]^{-1} \\ \Sigma_t^+ &= \Sigma_t - \Sigma_t F [F' \Sigma_t F]^{-1} F' \Sigma_t \end{aligned} \quad (35)$$

These update equations lead to a singular variance-covariance matrix Σ_t^+ . However $\hat{b}^+(t)$ and Σ_t^+ still define the density function of $b(t)$ on the domain in which $b(t)$ satisfies the natural equality constraints. Outside this domain the density function is zero. As a result Equation 30 transforms into:

$$\begin{aligned}
& \min b(t)[b(t) - \hat{b}^+(t)]', \sum_t^{+pinv} [b(t) - \hat{b}^+(t)], \\
& 0 \leq b_{(i-1)n+j}(t) \leq 1, i = 1 \dots m, j = 1 \dots n \\
& \sum_{j=1}^n b_{(i-1)n+j}(t) = 1, i = 1 \dots m
\end{aligned} \tag{36}$$

where pinv is defined as the pseudo-inverse operator [see also Anderson and Moore (18)].

EXPERIMENTS

Experiments were carried out with both simulated and real data. The advantage of using simulated data is that the original matrix is available to evaluate the different methods. However these experiments give only limited insight into whether a method would work in practice. Therefore a second series of experiments was done by using minute-by-minute induction loop data from the Amsterdam beltway.

First Experiment, Simulated Data

The simulated data have been obtained by programming the trip generation model shown in Figure 1(a). The on-ramp volumes were generated by using a Poisson random generator. The split probabilities were obtained by taking a weighted sum of two extreme split vectors:

$$\begin{aligned}
B(t) &= \alpha(t)B_1 + [1 - \alpha(t)]B_2, \\
\alpha(t) &= \frac{1}{2}[1 + \cos(2\pi t / T)], T = 144
\end{aligned} \tag{37}$$

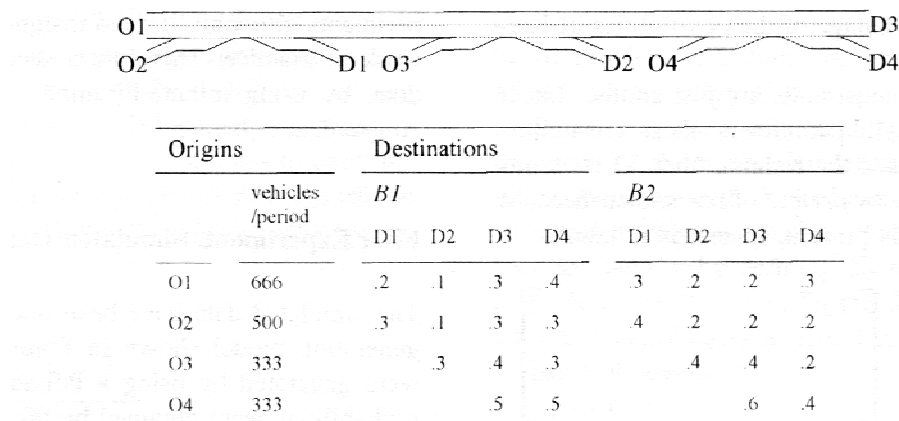
The network used consists of four entrances and four exits [Figure 3(a)].

As an evaluation criterion the square root of the mean squared error (RMSE) of the split parameters was used, that is,

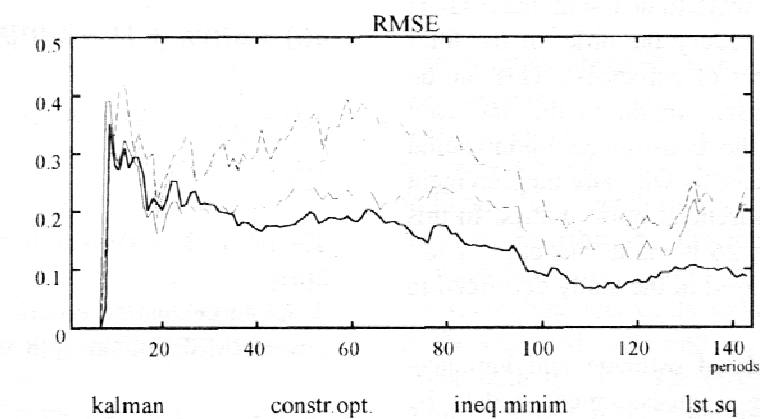
$$RMSE = \sqrt{\frac{1}{mn} \sum_{i=1}^m \sum_{j=1}^n [\hat{b}_{ij}(t) - b_{ij}(t)]^2} \tag{38}$$

All methods described in this paper were tested. To make a fair comparison all methods were optimized for parameters that reflect the rate of change in the dynamic OD. The results from the previous section were used to determine the noise error covariance matrix required by the Kalman filter and to the natural constraints. For this experiment the Kalman-based method

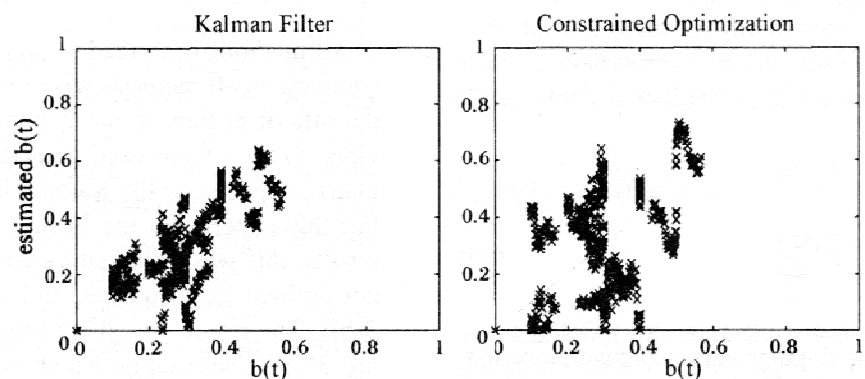
produced the best results; this was followed by constrained optimization, inequality constrained least squares, and ordinary least squares [see Figure 3(b)]. The results are also presented in scatter diagrams [see Figure 3(c)]. These diagrams show for a number of periods the estimated split values plotted against the real values.



(a)



(b)



(c)

FIGURE 3 Simulation results: (a) simulation setup, (b) simulation results, and (c) true versus estimated $b(t)$, $t > 100$.

Second Experiment, Empirical Data

The second series of experiments was done by using induction loop data from the Amsterdam beltway. For this experiment one direction of an 11-km freeway corridor was selected. This corridor has five entrances and five exits and is equipped with 19 detector stations. All data were aggregated to periods of 5 min. Again various methods were compared. This time only the diagonal elements of the variance covariance matrix prescribed by Equation 32 were used while applying the Kalman filter.

For this experiment observed trip matrices were not available. Therefore the evaluation criterion in Equation 38 could not be used. Instead the flow-predicting capabilities for a set L of reference locations were used. Set L is a set of n_L reference locations. It contains induction loops on locations for which the volumes are expected to be sensitive to the split parameters, for example, between off-ramps and on-ramps. To prevent data from being used at the same time to calculate and evaluate $\hat{b}(t)$, the volumes were predicted by multiplying 5-min-old split parameter estimates by

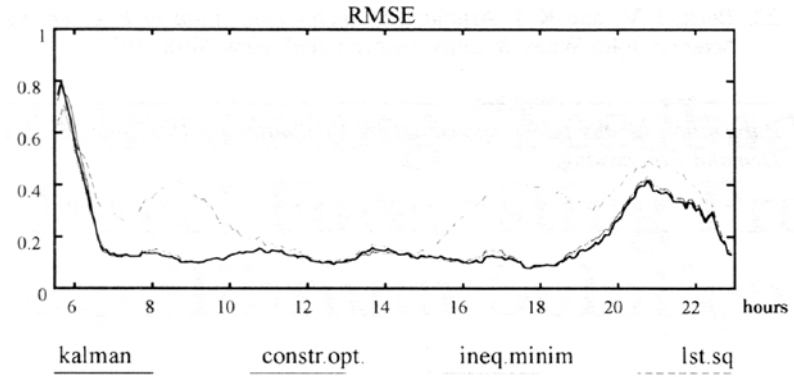
Equation 39 is not available.

With the evaluation criterion in Equation 39 it was not possible to prove major differences in performance between Kalman filtering, constrained optimization, and inequality-constrained least squares. Only the unconstrained least-squares method clearly gave results worse than those obtained by the other methods. After the evening rush hour the RMSE for all methods increased, probably because of suddenly changing OD patterns. When data from other days were evaluated RMSE plots with similar patterns appeared. This indicates that it might be useful to use a historic data base in which the permitted rate of change or even the direction of the changes in OD patterns are stored. Of all of the evaluated methods the Kalman filter seems the most suitable one for use in working with such a data base.

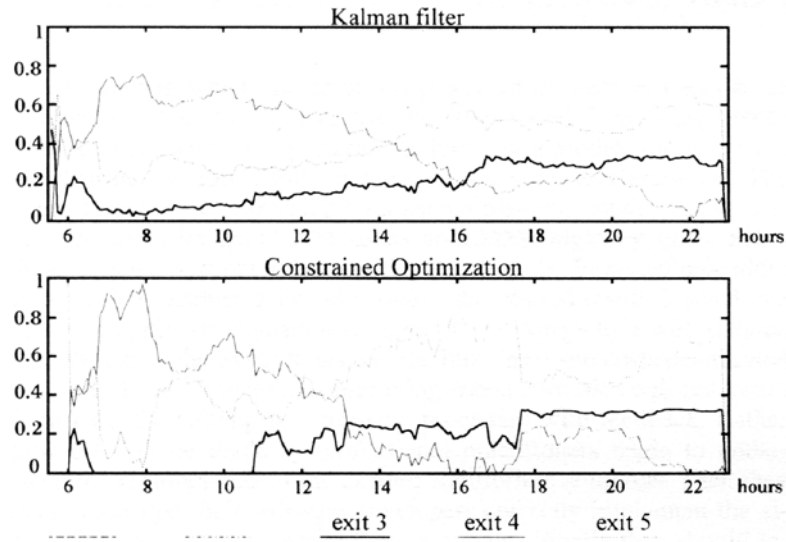
Although RMSE values do not differ significantly, comparing the split proportions estimated by different methods shows significant differences in estimated value; see for example Figure 4(b), which shows estimated splits for both the Kalman filter and the constrained least-squares methods.

To decide which of the two sets of parameters is more likely to correspond to the observed volumes, a second measure of effectiveness is introduced: the value of the likelihood function of the observations $y(t)$. Again $\hat{b}(t)$ is replaced by $\hat{b}(t-1)$ to prevent the use of observed volumes for estimation and evaluation purposes at the same time. The resulting likelihood is defined by

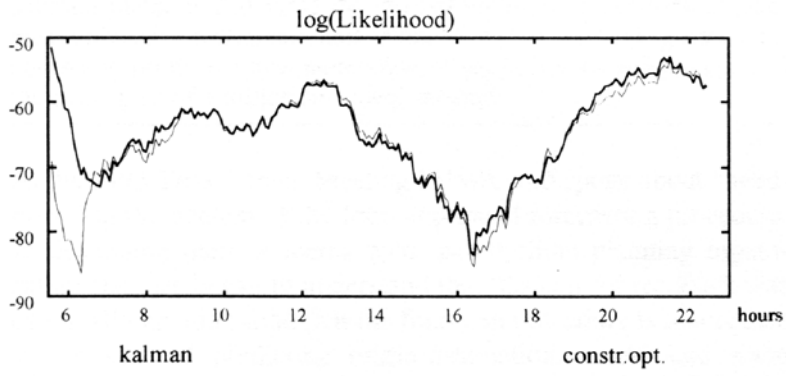
$$\begin{aligned} p[y(t)|\hat{b}(t)] &\approx \frac{1}{(2\pi)^{p/2} \sqrt{|R_{t-1}|}} \exp - \frac{1}{2} \\ &\times [y(t) - H'(t)\hat{b}(t-1)]' \\ &\times R_{t-1}^{-1} [y(t) - H'(t)\hat{b}(t-1)] \end{aligned} \quad (40)$$



(a)



(b)



(c)

FIGURE 4 Empirical results: (a) moving average of RMSE, (b) estimated splits, entrance 4, and (c) moving average of log (likelihood).

In Figure 4(c) the moving average of the logarithm of this likelihood is displayed. Figure 4(c) shows that a test of the hypothesis by using a likelihood ratio would generally favor the Kalman filter-generated solution.

CONCLUSIONS

The problem of estimating dynamic OD matrices was converted to the problem of estimating split parameters in a trip generation model. A Kalman-based method was compared with other methods like least squares and constrained optimization.

A new way of initializing the Kalman filter and of imposing the natural inequality and equality constraints was derived from theory. A measurement noise covariance matrix that was derived from the trip generation model was used.

The resulting method was programmed and tested. Tests with simulated data indicate that the Kalman-based filter method performs better than the other methods. Tests with real data indicate that results can be improved by using a Kalman filter combined with a data base in which optimal tuning parameters for the filter are stored.

REFERENCES

1. Cascetta, E., and S. Nguyen. A Unified Framework for Estimating or Updating Origin/Destination Matrices from Traffic Counts. *Transportation Research B*, Vol. 22B, No. 6, 1988, pp. 437-455.
2. Hamerslag, R., and B. H. Immers. Estimation of Trip Matrices: Shortcomings and Possibilities for Improvement. In *Transportation Research Record 1203*, TRB, National Research Council, Washington, D.C., 1988, pp. 27-39.
3. Bell, M. G. H. Variances and Covariances for Origin-Destination Flows When Estimated by Log-Linear Models. *Transportation Research-B*, Vol. 19B, No. 6, 1985, pp. 497-507.
4. Hendrickson, C., and S. McNeil. Estimation of Origin/Destination Matrices with Constrained Regression. In *Transportation Research Record 976*, TRB, National Research Council, Washington, D.C., 1984.
5. van Zuylen, H. J., and L. G. Willumsen. The Most Likely Trip Matrix Estimated from Traffic Counts. *Transportation Research-B*, Vol. 14B, 1980, pp. 281-293.
6. Cremer, M., and H. Keller. Dynamic Identification of Flows from Traffic Counts at Complex Intersections. *Proc., Eighth International Symposium on Transportation and Traffic Theory, 1981*.
7. Cremer, M. Determining the Time-Dependent Trip Distribution in a Complex Intersection for Traffic Responsive Control. *IFAC Control in Transportation Systems*. Baden-Baden, Germany, 1983.

8. Cremer, M., and H. Keller. A New Class of Dynamic Methods for the Identification of Origin-Destination Flows. *Transportation Research-B*, Vol. 21B, No. 2, 1987, pp. 117-132.
9. Nihan, N. L., and G. A. Davis. Recursive Estimation of Origin-Destination Matrices from Input/Output Counts. *Transportation Research-B*, Vol. 21B, No. 2, 1987, pp. 149-163.
10. Nihan, N. L., and G. A. Davis. Application of Prediction-Error Minimization and Maximum Likelihood to Estimate Intersection O-D Matrices from Traffic Counts. *Transportation Science*, Vol. 23, No. 2, May 1989.
11. Bell, M. G. H., D. Inaudi, J. Lange, and M. Maher. Techniques for the Dynamic Estimation of O-D Matrices in Traffic Networks. *Proc., Drive Conference*, February 4 to 6, 1991.
12. Keller, H., and G. Ploss. Real-Time Identification of O-D Network Flows from Counts for Urban Traffic Control. *Proc., 10th Symposium on Traffic Theory*, 1987.
13. Bell, M. G. H. The Real Time Estimation of Origin-Destination Flows in the Presence of Platoon Dispersion. *Transportation Research-B*, Vol. 25B, 1991, pp. 115-125.
14. van der Zijpp, N. J., and R. Hamerslag. The Real Time Estimation of Origin-Destination Matrices for Freeway Corridors. *Proc., 26th ISATA Conference*, 1993.
15. Davis, G. A., and N. L. Nihan. A Stochastic Process Approach to the Estimation of Origin Destination Parameters from Time-Series of Traffic Counts. Presented at 70th Annual Meeting of the Transportation Research Board, Washington, D.C., 1991.
16. Davis, G. A. *A Stochastic Dynamic Model of Traffic Generation and Its Application to the Maximum Likelihood Estimation of Origin Destination Parameters*. Ph.D. thesis. University of Washington, 1989.
17. Dempster, A. P., N. M. Laird, and D. B. Rubin. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society*, Vol. 39, Series B, 1977, pp. 1-38.
18. Anderson, B. D. O., and J. B. Moore. *Optimal Filtering*. Prentice-Hall, Incorporated, Englewood Cliffs, N.J., 1979.
19. Catlin, D. E. Estimation, Control, and the Discrete Kalman Filter. *Applied Mathematical Sciences*, 71. Springer-Verlag, 1989.
20. Maher, M. J. Inferences on Trip Matrices from Observations on Link Volumes: A Bayesian Statistical Approach. *Transportation Research B*, Vol. 17B, No. 6, 1983, pp. 435-447.
21. van der Zijpp, N. J., and R. Hamerslag. A Bayesian Approach to Estimate Origin-Destination Matrices for Freeway Corridors. Presented at the Universities Transport Study Group 1994 Conference, 1994.
22. Beck, J. V., and K. J. Arnold. *Parameter Estimation in Engineering Science*. John Wiley & Sons, Incorporated, New York, 1977.

Publication of this paper sponsored by Committee on Passenger Travel Demand Forecasting.

Introducing "Feedback" into Four-Step Travel Forecasting Procedure Versus Equilibrium Solution of Combined Model

DAVID E. BOYCE, YU-FANG ZHANG, AND MARY R. LUPA¹

The manner in which the solutions produced by various methods of introducing "feedback" into the four-step travel forecasting procedure compare with the equilibrium solution of a model combining the trip distribution, mode split, and assignment steps was examined. The comparisons were performed on a sketch-planning model of the Chicago region with about 300 zones and 3,000 highway links. From these comparisons one can learn that iterating the four-step procedure in an ad hoc manner does not produce the desired result. Instead one needs to apply an algorithm designed to converge to a well-defined equilibrium of the travel flows and the link times and costs determined by these flows. Progress in improving travel forecasts will not result from calls for solving the four-step procedure with feedback. Rather progress will be made as professional practitioners begin to understand the requirements of the desired equilibrium solutions. Then they must insist that their software developers correctly implement the algorithms required to compute these solutions. Finally they should insist that FHWA short courses introduce participants to contemporary solution methods that yield the desired equilibrium properties. Likewise university instructors and textbook authors should update their courses to produce a new generation of professionals who understand the principles of equilibrium travel models.

At the 1993 TRB Annual Meeting FHWA staff spoke about "feedback" in the context of the four-step travel forecasting procedure. In discussing their concerns with metropolitan planning organization staff we began to understand that this call for feedback was essentially an admission that the four-step procedure is inadequate to the task of predicting origin-destination, mode, and route choices in a congested, multimodal urban transportation network. Because they were not sure what to do about this inadequacy and because they were mired in the paradigm of the 1960s, they were calling for the solution of the 1960s: iterate the four-step procedure until the link flows, their associated generalized travel costs (impedances), and the corresponding origin-destination-mode choices are brought into a consistent relationship with each other.

The first author of this paper, having entered the urban transportation planning field in about 1960, remembers well the efforts of early models to define what they meant by feedback and convergence in the emerging four-step procedure. In the course of interviewing the staffs of various metropolitan planning agencies in 1968 for his book, *Metropolitan Plan Making (1)*, he recalls asking whether they had ever succeeded in iterating their travel forecasting procedure, that is, resolving the four-step procedure by using the travel times yielded by the trip assignment step. The answer was universally no. Neither they nor he had considered rough approach.

Since becoming aware of the formulations of Evans (2) and Florian et al. (3) in mid-1976, the first author rarely thought about feedback in the above sense until January 1993. Instead he has devoted the past 17 years to implementing, evaluating, and calibrating various models, mainly for

¹ D.E. Boyce and Y.-F. Zang, Urban Transportation Center, University of Illinois at Chicago, 1033 West Van Buren Street, Suite 700 South, Chicago, Ill. 60607. M.R. Lupa, Chicago Area Transportation Study, 300 West Adams Street, Chicago, Ill. 60606.

the Chicago region, which are guaranteed by their formulation and solution method to converge to the *equilibrium* solution that is still characterized by the obsolete term *feedback* [see Boyce et al. (4-6), Boyce and Lundqvist (7), Boyce (8), and Lee (9)]. Putman (10) has applied the same concept to a small test problem as well as to larger-scale problems.

Since it is apparent from the above remarks that the four-step procedure is finally viewed as inadequate, we thought it would be best to try to demonstrate what difference a convergent algorithm makes to the solution of travel choice models. By comparing those results with various approximate solutions used in practice, we hope to convince professional practitioners once and for all of the merits of the Evans partial linearization algorithm for solving combined models of trip distribution, mode choice, and assignment or, as we prefer to say, equilibrium models of origin-destination, mode, and route choice.

What are the characteristics of these equilibrium models that we find so appealing? In fact they are the same characteristics that we seek for the four-step procedure, but that are rarely seen in print, and were certainly not understood by the agency staff who originally proposed the four-step procedure in the 1960s [see for example Carroll (11)]. The two equilibrium conditions that we require may be simply stated as follows:

1. The generalized travel costs from each origin zone to each destination zone by automobile equal the sums of the individual link costs over the used routes; no unused route has a lower cost; the link costs depend in part on the link flows resulting from the trips per hour by automobile between all origin-destination pairs.
2. The number of trips per hour from each origin zone to each destination zone by each mode depends on the generalized travel costs determined in part by the automobile link flows resulting from those trips.

Perhaps Beckmann et al. (12) put it best, as well as first: "The prevailing demand for transportation, that is the existing pattern of origins and terminations, gives rise to traffic conditions that will maintain that same demand." Unfortunately almost no one in this field, including the first author, was aware of their fundamental contribution to our field when it was so greatly needed in the late 1950s.

The intended contribution of this paper is not to describe in detail the Evans partial linearization algorithm for solving equilibrium models, which produces solutions that are guaranteed to converge to the above conditions. That has been done elsewhere (5,6,13): however, the method is described below in terms familiar to practitioners. Rather the intent is to compare the results of this method with various solution techniques used at present and in the past that are *intended* to converge to the desired conditions. These techniques are a sampling of possible iterative approaches and are intended to be illustrative. The objective of all of the approaches is the same: to find the solution that satisfies the equilibrium conditions. Some methods can be proven to converge to these conditions: others cannot. The issue is which ones do converge and how quickly do they achieve an acceptable level of convergence?

Following a statement of four solution techniques, as well as the Evans algorithm, the results of solving a large-scale model with each method are compared. The variables used in this comparison are highway link flows, automobile and transit trip tables, and automobile generalized costs: transit generalized costs are a fixed input to all the methods and hence do not

vary. The solution variables are compared with a highly converged solution of the model, which may be regarded as the "true" solution. Such a highly converged solution would not usually be computed in practice, and hence serves as a standard for comparing the various methods.

The paper concludes with the authors' recommendations concerning what steps should be taken to implement the use of equilibrium models in professional practice.

COMPARISON OF SOLUTION METHODS

In this section we describe the five solution procedures applied in our experiments. First, we describe the variants of procedures based on traditional practice. Four procedures were defined. In each procedure an estimate of the automobile generalized travel cost matrix and the fixed transit generalized cost matrix are inputs to the trip distribution model. Following the Chicago Area Transportation Study practice of using an automobile travel cost matrix for the most similar network available, we used the matrix from the fifth iteration of the Evans algorithm described below. This choice gives each procedure a highly advantageous starting point.

Method 1: One Iteration of Trip Distribution, Mode Split and AON Assignment

We begin with the simplest possible choice, one iteration of the four-step procedure with all-or-nothing (AON) assignment as the assignment method. Although this procedure is believed to be completely inadequate, it does provide one simple result for comparison with other methods. The procedure is illustrated in Figure 1(a) with one iteration.

Method 2: Multiple Iterations of Trip Distribution, Mode Split and AON Assignment

This procedure is the simplest concept of feedback: simply iterate through the four-step procedure several times; the travel costs determined by the assignment step form the basis for the next trip distribution and mode choice. This procedure is also believed to be unsuitable; again we include it for comparative purposes. Method 2 is also illustrated by Figure 1(a).

Method 3: Multiple Iterations of Trip Distribution, Mode Split, and AON Assignment with Averaging at Each Iteration

This procedure is similar to the previous one except that the origin-destination-mode matrix and the link flow vector are averaged together after each solution of the four-step procedure. The weights are chosen as follows:

1. The first solution results from one iteration of the four-step procedure (same as Method 1 above).
2. The second solution, which is based on the travel costs of the first solution, is averaged with the first solution with equal weights (50/50).
3. The third solution, which is based on the average of the first two solutions, is weighted one-third and the former solution is weighted two-thirds (67/33).
- n. The nth solution, which is based on the result of iteration ($n - 1$), is weighted ($1/n$) and the

former solution is weighted $(n - 1)/n$.

Note that in each solution each of the previous solutions is weighted equally; moreover at each iteration the combined results from the previous solutions provide the inputs to the four-step procedure. This method is somewhat like the Bureau of Public Roads capacity-restrained assignment; in that procedure, however, the link-flow vectors were averaged only as the conclusion of several iterations.

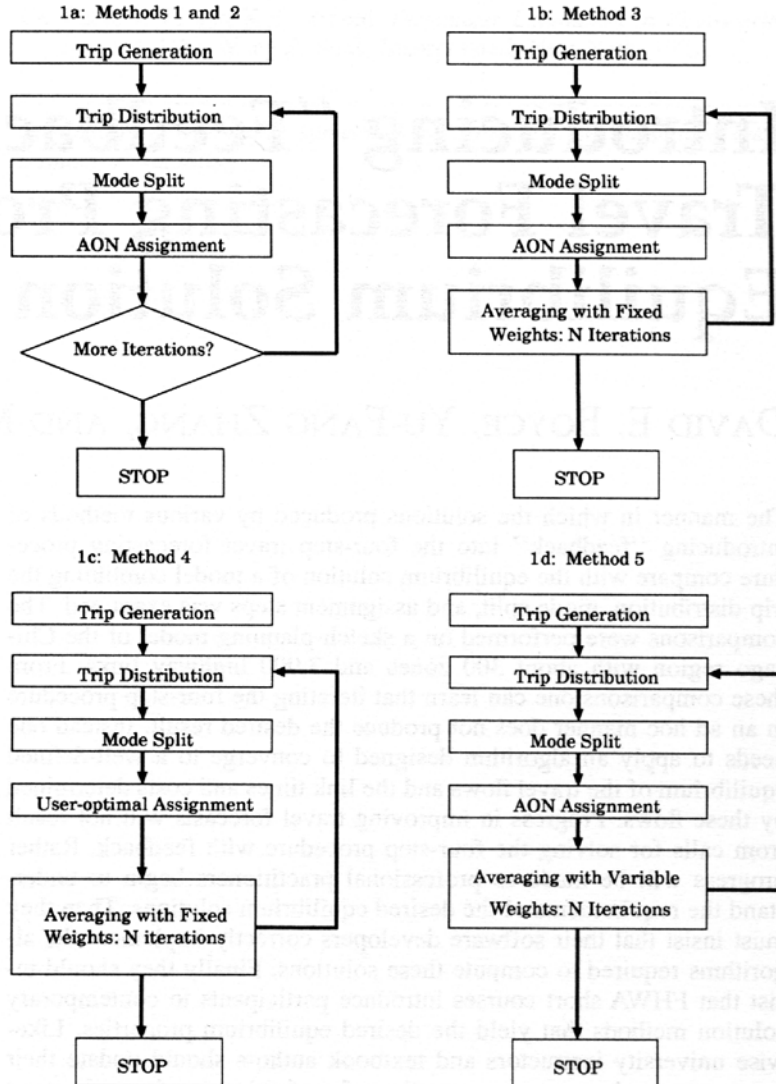


FIGURE 1 Comparison of solution procedures: (a) Methods 1 and 2, (b) Method 3, (c) Method 4, and (d) Method 5.

Method 3 is known in the transportation science literature as an iterative technique that uses

predetermined step sizes or the method of successive averages. Under the conditions that are satisfied here, the method is known to converge to the desired equilibrium solution [see Sheffi (14,p. 324)]. However convergence may be quite slow. Method 3 is shown in Figure 1(b).

Method 4: Multiple Iterations of Trip Distribution, Mode Split, and User-Optimal Assignment with Averaging at Each Iteration

This method is similar to Method 3 except that the AON assignment is replaced by a user-optimal assignment with five iterations. User-optimal assignment, as performed by the Frank-Wolfe or linearization method, consists of the following steps:

1. Perform an AON assignment of the automobile trip matrix to the automobile network;
2. In the first iteration designate the resulting link-flow vector as the current solution and return to step 1; in the second and successive iterations determine a weight for averaging the AON link-flow vector with the current solution (weighted average of previous AON link-flow vectors) such that the resulting vector is as close as possible to the user-optimal conditions for a fixed automobile trip matrix, as judged by a function of the new current solution; and
3. Check convergence and continue if the algorithm has not adequately converged; in this application the procedure was terminated after five iterations (AON assignments and averaging steps).

Feedback can be introduced by repeating this four-step procedure a second, third or fourth time. To ensure convergence the results of these iterations should be averaged together by applying the method of successive averages described under Method 3. Otherwise the results tend to oscillate.

Since each sequence of this four-step procedure involves five AON assignments, the computations for one iteration of Method 4 are roughly comparable to five iterations of Methods 2, 3, or 5. Method 4 is illustrated in Figure 1(c).

The four methods described above are intended to represent methods used in conventional practice. Next we turn to a description of an efficient, convergent algorithm for solving the equilibrium problem described earlier. At this point it may be helpful for the reader to think of the problem to be solved as the *underlying* equilibrium problem and to regard the traditional four-step procedure as a relatively crude solution method, or heuristic. In other words the problem we are seeking to solve is not some embellished version of the four-step procedure; rather we seek a solution that satisfies the two conditions stated earlier.

Method 5: Evans Algorithm

The algorithm described below is a partial linearization method, in contrast to the full linearization, or Frank-Wolfe, method mentioned above. We also refer to it as the *Evans algorithm* after its originator, Suzanne P. Evans (2). The method may be described informally as follows:

1. Solve the trip distribution and mode split steps of the four-step procedure, given an initial

automobile travel cost matrix, as well as the fixed transit cost matrix;

2. Perform an AON assignment of the automobile trip matrix to the automobile network;
3. In the first iteration, designate the trip matrices and link flow vector from Steps 1 and 2 as the current solution and return to Step 1; in the second and successive iterations determine a weight for averaging the trip matrices and the link-flow vector from Steps 1 and 2 with the current solution, such that the resulting matrices and vector are as close as possible to the equilibrium conditions described in the first section, as judged by a function of the new current solution (trip matrices and link-flow vector); and
4. Check convergence by a measure defined on the current solution and the results of Steps 1 and 2 and continue if the algorithm has not adequately converged.

The above method is the same as Method 3 except that the weights used in the averaging steps are not predetermined but are chosen to be the best at each iteration. A well-defined convergence measure is available on the basis of the calculation of the greatest lower bound on the objective function of the equivalent optimization problem at each iteration. This measure is very useful in monitoring the convergence and in comparing the convergence of solutions of alternative plans [see Figure 1(d)].

In the results that follow the solutions of the model computed with the Evans algorithm with 5, 10, 15, and 20 iterations are compared with those computed by Methods 1 to 4. Each of the solutions is compared with a "true" solution of the model computed with the Evans algorithm with 50 iterations. In this solution the objective function is no more than 0.2 percent from the optimal value desired for all methods, a very highly converged solution. Of course such a solution would not be utilized in practice; it is used here only to provide a basis for evaluating all of the methods. The "true" solution could also have been computed by Method 3 or 4 or any convergent method; we used the Evans algorithm because it is more efficient than Method 3 or 4.

Each of the methods described above was solved for a sketch-planning model of the six-county Chicago region. The model is based on a highly aggregated zone system, with 317 zones of 14.5 and 58 km² (9 and 36 Mi² each and about 3,000 highway links. The trip distribution and mode split model are a single exponential (logit) function doubly constrained to satisfy fixed trip ends. All trip purposes are combined in the single function. Transit choice is based on fixed matrices of transit in-vehicle times, waiting and transfer times, plus a transit fare matrix. The submodes of bus, rapid transit, and commuter rail are represented in one matrix generated from a single transit network. Automobile person trips are converted to automobile vehicle trips on the basis of an average occupancy factor and are assigned to the aggregated automobile network; a separate matrix of equivalent truck trips is also assigned. The generalized cost and trip deterrence parameters were calibrated from matrices based on the 1980 Census and survey data. The generalized cost parameters include a transit bias term that results in an accurate prediction of regional mode split. The predicted mode split for transit trips to the central business district is within 5 percent of the observed value. Additional details and sensitivity analyses may be found in Boyce et.al. (6).

ANALYSIS OF RESULTS

In this section we present the results of the solutions of the five methods described previously. The results are presented in the form of a table for each of the four variables (link flow, auto and transit trips, and auto generalized costs). Each table compares the solution for each method (M) with the highly converged or true solution (1) by using the following measures:

$$\text{Root mean square error (RMSE): } \left[\sum_i^m (M_i - T_i)^2 / m \right]^{1/2} \quad (1)$$

$$\text{Chi - square: } \sum_i^m \left[\frac{(M_i - T_i)^2}{T_i} \right] \quad (2)$$

With R^2 for a regression with M as the dependent variable and T as the independent variable.

The data elements are the pairs of origin-destination-mode combinations or the pairs of links; m is the number of data elements with positive values. Zero values in the solutions were removed, since these values are a property of the model formulation or the data rather than the solution method.

For these measures the desired results are as follows:

1. The values of RMSE and chi-square should be zero;
2. The value of R^2 should be 1.

In each table the results for Method 1 and Methods 2 to 5 for five iterations are presented first. Then the results for Methods 3, 4, and 5 for 10, 15, and 20 iterations are compared. Note that the result of Method 1 is also the initial solution for Methods 2 to 5, and hence serves as a basis for comparing Methods 2 to 5 after five iterations.

Next the results are illustrated with two sets of four plots each for link flows and one set each for the other variables. The first set of link-flow plots corresponds to Methods 2 to 5; the second set is for Method 5 with 5, 10, 15, and 20 iterations. Link flows are examined in more detail because this variable is the slowest to converge.

Results for Automobile Link Flows

The first five rows of Table 1 show the results for the five methods with five iterations except for Method 1. The results for the first two methods are clearly unacceptable. Methods 3, 4, and 5 are rather similar; recall that Method 3 consists of five iterations of trip distribution, mode split, and AON assignment with predetermined weights, Method 4 is one solution of the trip matrices and five iterations of user-optimal assignment, and Method 5 consists of five solutions of the trip matrices and five AON assignments. Although the results of Methods 3 and 5 have better RMSEs, Method 4 has a better chi-square value. Hence the results of these three methods are quite similar. Both RMSE and chi-square are very effective in comparing the solutions with the

true solution; however, R^2 is largely ineffective except for Methods 1 to 2, which are clearly very poor.

TABLE 1 Results of Five Methods for Automobile Link Flows

Method	Iteration	RMSE	Chi-sq	R-sq
1-5	1	2052	1884910	.56
2	5	5304	10199398	.44
3	5	386	464700	.84
4	1×5	428	230588	.92
5	5	358	504876	.98
3	10	194	116300	.98
4	2×5	333	196623	.98
5	10	310	87004	.98
3	15	125	49530	.99
4	3×5	293	181810	.98
5	15	145	21952	.99
3	20	85	26240	.99
4	4×5	244	146877	.98
5	20	77	7461	.99

number of links with positive flow - 2,767

Rows 6 to 8 of Table 1 compare two iterations of Method 4 with 10 iterations of Methods 3 and 5. Since these three methods involve 10 AON assignments, the computational effort is roughly comparable. Table 1 shows that Method 4 (2×5 iterations) converges only slightly from the 1×5 solution, whereas Methods 3 and 5 (10 iterations) continue their convergence. The RMSE and chi-square values for the Evans algorithm decrease more than for Methods 3 and 4 as the number of iterations increases (see rows 9 to 14 of Table 1). The convergence is more pronounced for chi-square than for RMSE.

Taming to Figures 2 and 3 one can observe that Method 2 produces unacceptable results by comparing the link flows for each method on the Y-axis with the "true" values on the x-axis. The results for Methods 3, 4, 5 are much closer to the 45-degree line, although Method 3 is much more dispersed than Methods 4 and 5. These results illustrate why scatter diagrams are essential in addition to measures such as RMSE and chi-square. Figure 3 shows the plots for higher numbers of iterations of the Evans algorithm (Method 5). The clustering around the 45-degree

line becomes more and more pronounced as the number of iterations increases.

Results for Automobile Trips

For the automobile and transit trip matrices the results for Methods 1 and 4 (1×5) are the same, since both are based on the initial matrix of generalized automobile costs. Both sets of measures are rather large in comparison with Methods 3 and 5, which involve solving the trip matrices five times. In Table 2 Method 2 is seen to be unacceptable, but Method 3, the predetermined step size method, is relatively good but is always inferior to Method 5. Examination of Method 4 shows that the method converges, but not as rapidly as Methods 3 and 5. Figure 4 also shows that Methods 2 and 4 are inferior to Methods 3 and 5.

Results for Transit Trips

Since transit trips are based on fixed generalized costs, in our model they depend only indirectly on the equilibrium automobile costs. Nevertheless, Methods 3 and 5 clearly produce superior results compared with those produced by Methods 1, 2, and 4 (see Table 3). Method 4 does converge, but more slowly than Methods 3 and 5. Figure 5 shows that Method 5 produces slightly better results than Methods 3 and 4; as before Method 2 is clearly unacceptable. Note that here, as with automobile trips, the plot for Method 1 would be identical to the plot for Method 4.

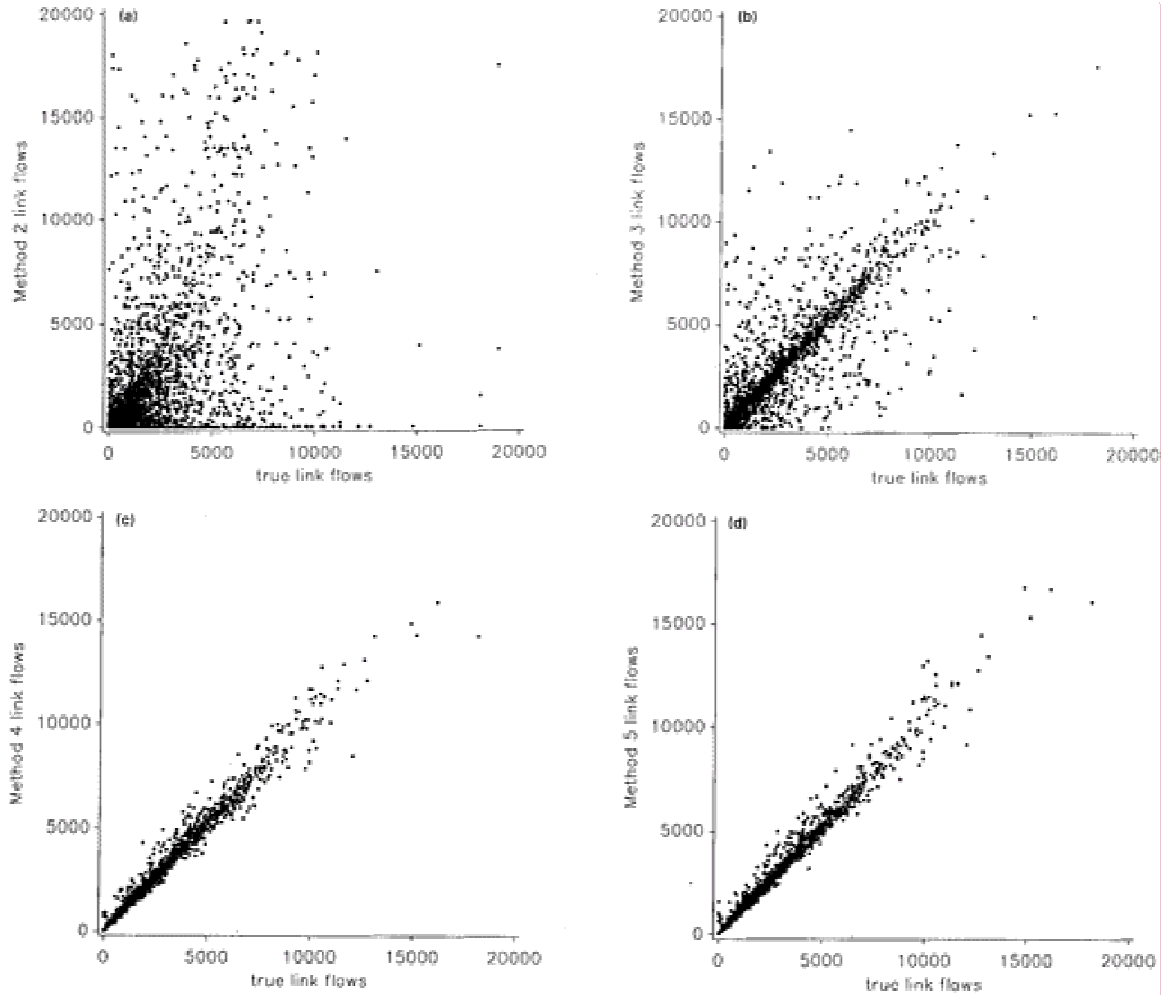


FIGURE 2 Results for link flows (Methods 2 to 5): (a) Method 2, (b) Method 3, (c) Method 4, and (d) Method 5.

Results for Automobile Generalized Costs

Table 4 for automobile generalized costs shows a pattern similar to those in Tables 2 and 3. Methods 1 and 2 are clearly inferior. Method 4 (1×5 iterations) is the next largest; this is followed by Methods 3 and 5. Method 4 converges, but not nearly as much as Methods 3 and 5 for 10, 15, and 20 iterations. The plots for these solutions, shown in Figure 6, are rather similar for Methods 3, 4, and 5, which show much better convergence than Method 2.

Results for Selected Regional Attributes

One important question raised about an earlier version of this paper is, "Does the choice of method make a difference in an important model output variable like vehicle kilometers of travel (VKT)?" We had to admit that we did not know the answer to this important question, but we decided to find out. The results are presented in Table 5.

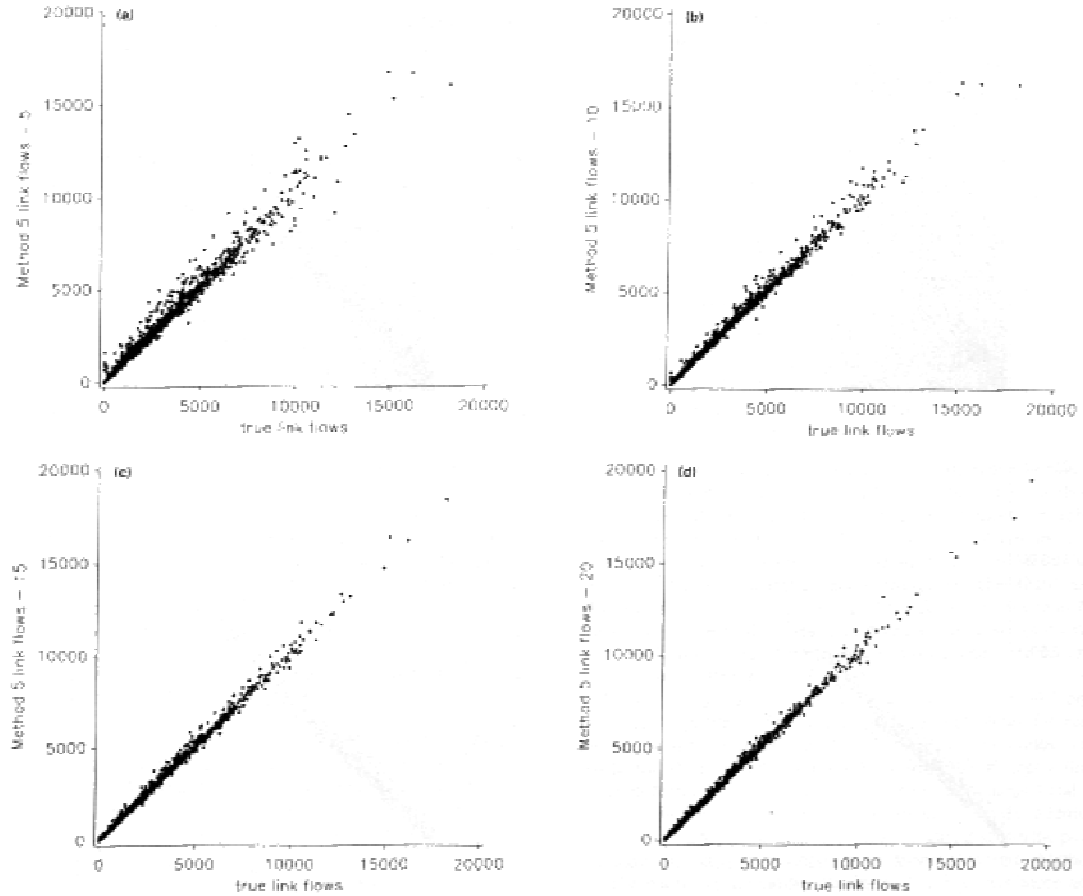


FIGURE 3 Results for link flows (Method 5): (a) 5 iterations, (b) 10 iterations, (c) 15 iterations, and (d) 20 iterations.

From several regional attributes computed by our code we selected highway vehicle kilometers of travel, mean automobile travel time, automobile space-mean-speed, and percentage of trips by transit. Central processing unit (CPU) time is also included. The results were quite surprising to us, and therefore well worth presenting. We know from Tables 1 to 4 that the choice of method does lead to important differences in measures comparing differences in the model outputs at the link or zone pair level. At the regional level, however, the aggregated model attributes are essentially the same for Methods 2, 4, and 5. Method 2, which is not a convergent method, yields unacceptable values however. We also observed, but do not report here, that Method 4 does diverge if the averaging step is omitted.

TABLE 2 Results of Five Methods for Automobile Trips

Method	Iteration	RMSE	Chi-sq	R-sq
1-5	1	7.38	9352	.99
2	5	28.25	338734	.86
3	5	4.48	4510	.99
4	1×5	7.38	9352	.99
5	5	3.78	3015	.99
3	10	4.48	1629	.99
4	2×5	3.94	3677	.99
5	10	1.80	645	.99
3	15	1.85	746	.99
4	3×5	3.30	3039	.99
5	15	1.25	292	.99
3	20	1.40	416	.99
4	4×5	3.11	2632	.99
5	20	.84	147	.99

number of positive O-D flows - 72,630

What those results indicate, then, is that any method that can be shown to converge to the true solution should yield reasonably good values of regional attributes. Recall, however, that there is no established concept of *convergence* as a requirement for the four-step procedure. Hence “Use a convergence method” is the proper response to the question, “How do I introduce feedback into the four-step procedure?”

Although the differences are small, the reader may notice that Method 5 is slightly superior to Methods 3 and 4 for 10 or more iterations for the regional attributes presented. The additional computing effort needed to obtain this result is a one-third increase over Method 4.

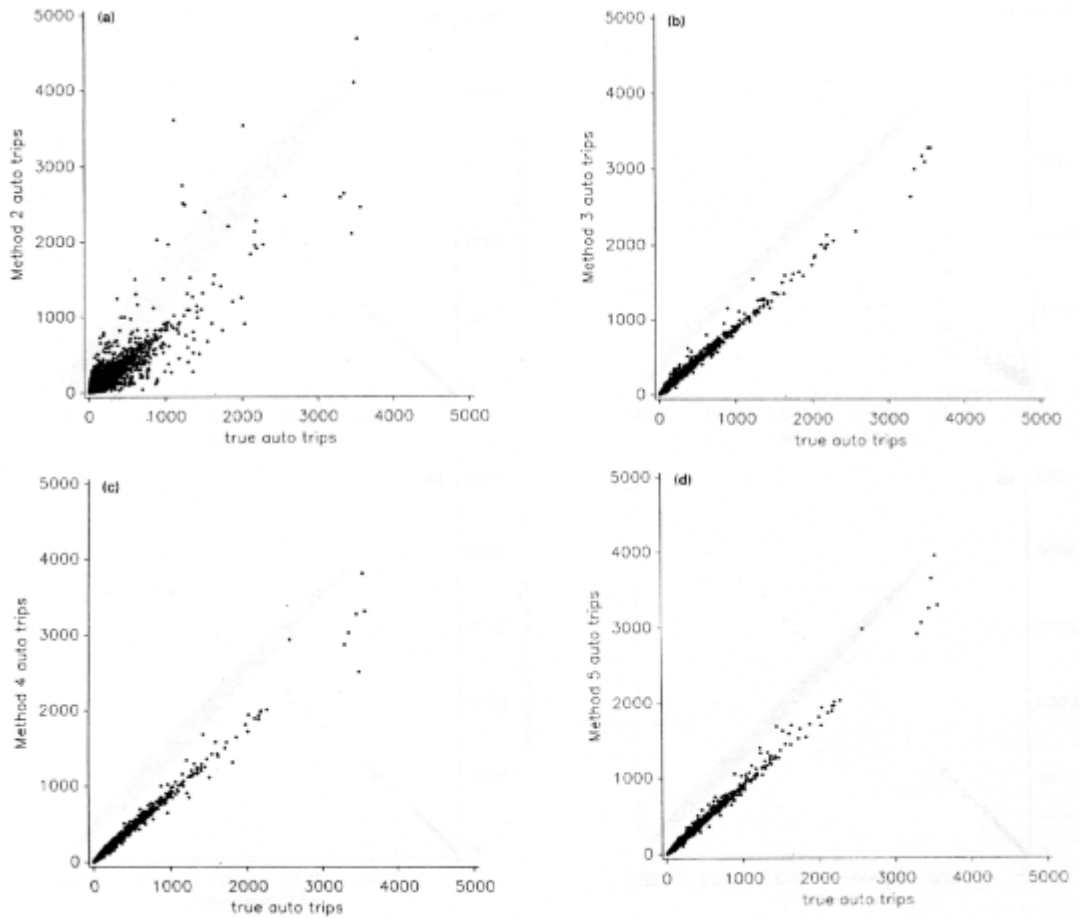


FIGURE 4 Results for automobile trips (Methods 2 to 5): (a) Method 2, (b) Method 3, (c) Method 4, and (d) Method 5.

TABLE 3 Results of Five Methods for Transit Trips

Method	Iteration	RM E	Chi-sq	R-sq
1-5	1	4.28	934	.99
2	5	25.39	22992	.96
3	5	1.23	92	.99
4	1×5	4.28	934	.99
5	5	1.05	52	.99
3	10	.79	37	.99
4	2×5	.81	58	.99
5	10	.58	19	.99
3	15	.52	18	.99
4	3×5	.76	45	.99
5	15	.47	10	.99
3	20	.38	11	.99
4	4×5	.82	39	.99
5	20	.36	6	.99

number of positive O-D flows - 55,141

CONCLUSIONS

Although it is recognized that these results are quite aggregated, we hope that they provide substantial insights into the performance of various methods, both ad hoc and convergent, for solving the travel forecasting procedure in an iterated manner. Although this is not the place for mathematical justifications of the Evans algorithm, we trust that the computational results produced by Method 5 are also convincing because they do converge to the desired equilibrium. What is equally important is that the computational effort for Method 5 is similar to those for Methods 2 and 3 and only slightly more onerous than that for Method 4. A time-saving variant of Method 5 is to update the trip matrices at every third or fifth iteration rather than at every iteration. The total number of iterations required for method 3, 4, or 5 depends on the desired convergence. At least five iterations are necessary for congested networks.

Rapid improvements in desktop computing speed and memory should continue to facilitate the use of more appropriate methods than in the past. As recently as 1987 the only computer available to us to solve our Chicago region model repeatedly was a Cray supercomputer. Now we are solving it routinely on a Sun SPARC-station 2 with 32 megabytes of memory in 3.1 min per

iteration. We believe that UNIX workstations of this type will be the computing platform of choice for planning agencies in the near future.

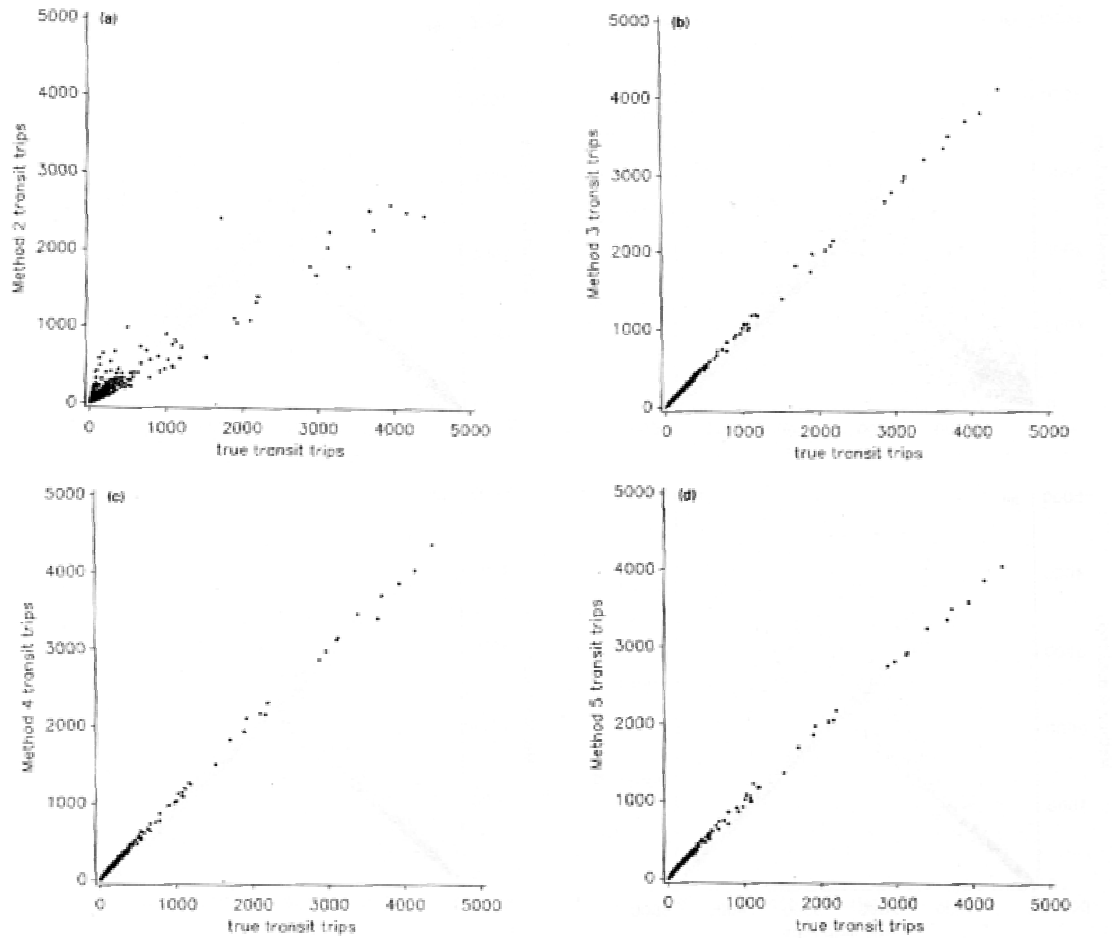


FIGURE 5 Results for transit trips (Methods 2 to 5): (a) Method 2, (b) Method 3, (c) Method 4, and (d) Method 5.

TABLE 4 Results of Five Methods for Automobile Costs

Method	Iteration	RMSE	Chi-sq	R-sq
1-5	1	.400	1877	.99
2	5	.664	5444	.97
3	5	.167	312	.99
4	1×5	.194	456	.99
5	5	.141	238	.99

3	10	.031	13	.99
4	2×5	.143	230	.99
5	10	.080	80	.99
3	15	.026	8	.99
4	3×5	.103	120	.99
5	15	.047	28	.99
3	20	.014	3	.99
4	4×5	.083	76	.99
5	20	.018	4	.99
number of O-D pairs -			100,489	

We conclude the paper with several observations concerning the implementation and adoption of equilibrium travel forecasting models. First, one might ask, why has it taken so long for convergent algorithms such as Methods 3 and 5 to be adopted in professional practice? We believe there are two aspects to this question. The first is that planning agencies apply models that are substantially more detailed than our own combined model. In particular they disaggregate origin-destination (O-D) tables by trip purpose and mode choice by user classes. Although these important disaggregations are not included in our present model, we are confident that it can be disaggregated in a similar manner. Moreover planning agencies have not faced much pressure to improve their methods until recently. In this situation they have continued to do what they knew from their own experience, which is the traditional four-step procedure.

The second aspect is that even if they had wished to adopt the new convergent methods the software was not available. "Why not?" one might reasonably ask, since some of the software developers are also leading researchers in transportation science. We asked this question on numerous occasions. The answer has been consistent: "software developers provide what the agencies demand."

We suggest that it is now time for metropolitan planning agencies, as well as FHWA and FTA, to demand software that yields the equilibrium solutions needed to meet the planning requirements of the 1990s. Let there be no further excuses on the part of practitioners and software developers for using obsolete methods.

Our next observation concerns the role of FHWA and FTA in training practitioners in their short courses. Is it not clear that the four-step procedure taught in these courses should be replaced by a modern approach? Again this will happen if the planning agencies demand it and not just accept at face value the recent statements about the need for "feedback." Surely they deserve better than this.

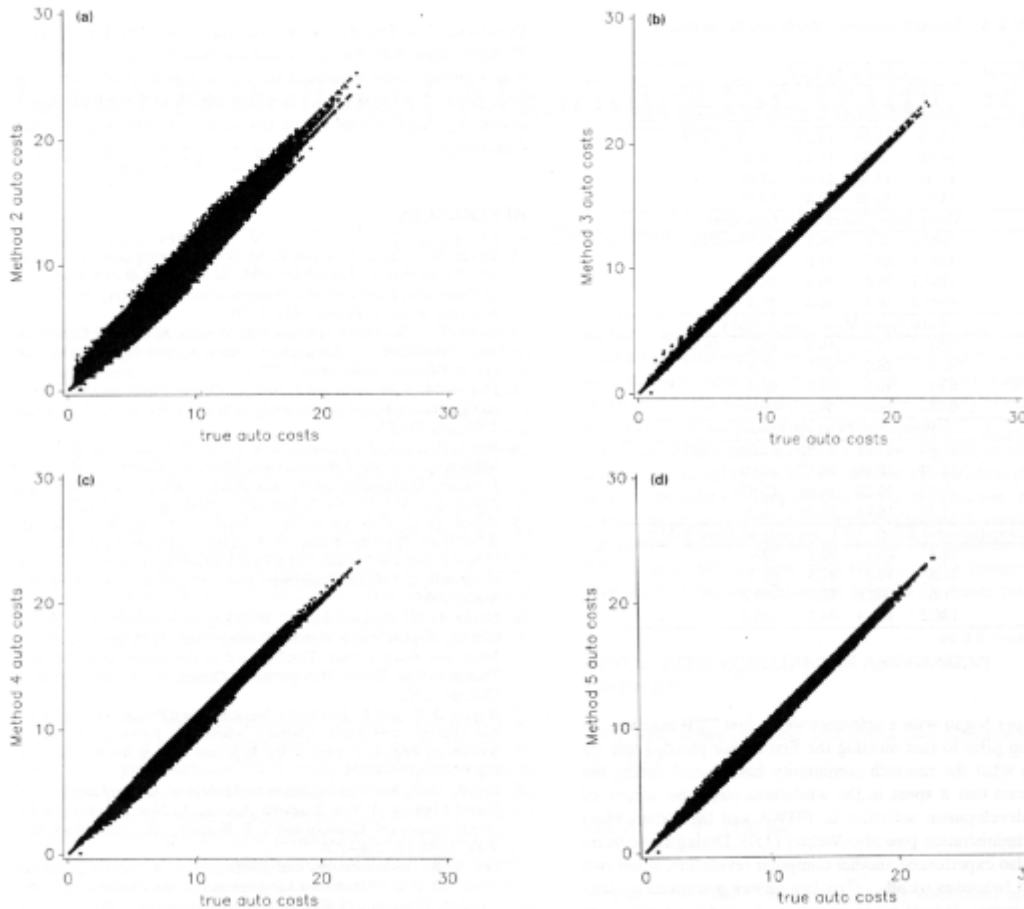


FIGURE 6 Results for automobile generalized costs (Methods 2 to 5): (a) Method 2, (b) Method 3, (c) Method 4, and (d) Method 5.

Finally those of us in the academic community need to assess our own shortcomings and responsibilities for this situation. We know of no book on travel forecasting methods that is accessible to professionals and describes a modern approach. Sheffi (14) has made a fine contribution to this subject, but his book is undoubtedly inaccessible to many practitioners and is now out of print. If we are to declare the four-step procedure obsolete there is no better place to begin than by writing new textbooks for both undergraduate and graduate courses. Such an effort will be substantial and would benefit from the financial support of FHWA and FFA under the Intermodal Surface Transportation Efficiency Act and the Clean Air Act Amendments.

TABLE 5 Selected Regional Attributes for Methods 2 to 5

Attribute	Solution Method				
Iterations	2	3	4	5	true
Highway Vehicle Kilometers of Travel (kilometers x 10 ⁶)					
5	17.19	13.56	13.17	13.43	13.23
10	14.06	13.39	13.44	13.30	
15	17.31	13.33	13.44	13.28	
20	14.07	13.30	13.41	13.27	
Mean Auto Travel Time (minutes)					
5	2037.9	27.6	26.7	26.9	26.3
10	4420.1	26.7	27.0	26.8	
15	1794.1	26.6	26.9	26.4	
20	4597.3	26.5	26.8	26.4	
Auto Space-Mean-Speed (kph)					
5	0.64	45.2	46.2	46.2	46.9
10	0.16	46.5	46.0	45.9	
15	0.64	46.5	46.2	46.7	
20	0.16	46.7	46.4	46.7	
Percent of Trips by Transit					
5	11-93	15.94	6.85	16.05	15.99
10	15.32	16-00	16.07	16.00	
15	12.02	16.20	16.01	15.97	
20	15.27	16.03	16.03	15.97	
Computational Effort (CPU seconds on Cray Y-MP)					
5	39.4	39.1	30.2	39.0	
10	51.8	70.3	51.5	69.5	
15	71.6	100.8	72.7	102.9	
20	136.2	131.4	94.1	132.4	

1 km = 0.6 mi.

This paper began with a reference to the last TRB meeting. At a workshop prior to that meeting the first author pleaded that we not forget what the research community has learned during the past 12 years that it spent in the wilderness since the demise of software development activities in FHWA and the Urban Mass Transit Administration [see also Weiner (15)]. During those

years we have also experienced another computer revolution, if not two or three. It behooves us all-planning agency practitioners, software developers, federal program managers, and academics-to work together to ensure that the next generations of travel forecasting methods benefit rapidly from research findings, practical experience, and advances in computing technology.

ACKNOWLEDGMENTS

An earlier version of this paper was presented at the Fourth National Conference on Transportation Planning Methods Applications in Daytona Beach, Florida. We are grateful for numerous comments they received on that paper. The sketch-planning model and solution algorithm, on which the results presented in the paper are based, were developed in collaboration with Dean B. Englund and Ronald W. Eash of the Chicago Area Transportation Study. We are grateful for their critical advice and encouragement. Financial support for this work was provided recently by the Illinois

Department of Transportation and earlier by the Urban Mass Transportation Administration and the National Science Foundation. Computational aspects of the research were partly supported by a grant of computer time from the National Center for Supercomputing Applications at the University of Illinois at Urbana-Champaign.

REFERENCES

1. Boyce, D. E., N. D. Day, and C. McDonald. *Metropolitan Plan Making: An Analysis of Experience with the Preparation and Evaluation of Alternative Land Use and Transportation Plans*. Regional Science Research Institute, Philadelphia, 1970.
2. Evans, S. P. Derivation and Analysis of Some Models for Combining Trip Distribution and Assignment. *Transportation Research*, Vol. 10, No. 1, February 1976, pp. 37-57.
3. Florian, M., S. Nguyen, and J. Ferland. On the Combined Distribution-Assignment of Traffic. *Transportation Science*, Vol. 9, No. 1, February 1975, pp. 43-53.
4. Boyce, D. E., L. J. LeBlanc, K. S. Chon, Y. J. Lee, and K. T. Lin. Implementation and Computational Issues for Combined Models of Location, Destination, Mode and Route Choice. *Environment and Planning A*, Vol. 15, No. 9, September 1983, pp. 1219-1230.
5. Boyce, D. E., K. S. Chon, M. E. Ferris, Y. J. Lee, K. T. Lin, and R.W. Eash. Implementation and Evaluation of Combined Models of Urban Travel and Location on a Sketch Planning Network. University of Illinois at Urbana-Champaign and Chicago Area Transportation Study, 1985.
6. Boyce, D. E., M. Tatineni, and Y. Zhang. *Scenario Analyses for the Chicago Region with a Sketch Planning Model of Origin-Destination, Mode and Route Choice*. Final Report to the Illinois Department of Transportation. Urban Transportation Center, University of Illinois, Chicago, 1992.
7. Boyce, D. E., and L. Lundqvist. Network Equilibrium Models of Urban Location and

- Travel Choices: Alternative Formulations for the Stockholm Region. *Papers of the Regional Science Association*, Vol. 61, 1987, pp. 9I-I04.
8. Boyce, D. E. Network Equilibrium Models of Urban Location and Travel Choices: A New Research Agenda. In *New Frontiers in Regional Science* (M. Chatterji and R. E. Kuenne, eds.). Macmillan, New York, 1990, pp. 238-2-56.
 9. Lee, C. K. *Implementation and Evaluation of Network Equilibrium Models of Urban Residential Location and Travel Choices*. Ph.D. dissertation. University of Illinois, Urbana-Champaign, 1987.
 10. Putman, S. H. *Integrated Urban Models 2*. Pion, London, 1991.
 11. Carroll, J. D., Jr. A Method of Traffic Assignment to an Urban Network. In *Highway Research Bulletin 224*, HRB, National Research Council, Washington, D.C., 1959, pp. 64-71.
 12. Beckmann, M., C. B. McGuire, and C. B. Winsten. *Studies in the Economics of Transportation*. Yale University Press, New Haven, Conn., 1956.
 13. Boyce, D. E., L. J. LeBlanc, and K. S. Chon. Network Equilibrium Models of Urban Location and Travel Choices: A Retrospective Survey. *Journal of Regional Science*, Vol. 28, No. 2, May 1988, pp. 159-183.
 14. Sheffi, Y *Urban Transportation Networks: Equilibrium Analysis with Mathematical Programming Methods..* Prentice-Hall, Incorporated, Englewood Cliffs, N.J., 1985.
 15. Weiner, E. Upgrading U.S. Travel Demand Forecasting Capabilities. *Urban Transportation Monitor*, Vol. 16. Lawley Publications, Burke, Va., 1993, p. 2.

Publication of this paper is sponsored by Committee on Transportation Supply Analysis.

Faster Path-Based Algorithm for Traffic Assignment

R. JAYAKRISHNAN, WEI K. TSAI, JOSEPH N. PRASHKER, AND SUBODH RAJADHYAKSHA¹

A fresh look at the arguments against path-enumeration algorithms for the traffic assignment problem is taken, and the results of a gradient projection method are provided. The motivation behind the research is the orders of magnitude improvement in the availability of computer storage over the last decade. Faster assignment algorithms are necessary for real-time traffic assignment under several of the proposed advanced traffic management system strategies, and path-based solutions are preferred. The results show that gradient projection converges in one-tenth of the iterations of the conventional Frank-Wolfe algorithm. The computation time improvement is of the same order for small networks but is reduced as the network size increases. The computer implementation issues are discussed carefully, and schemes to achieve a 10-fold speedup for larger networks are also provided. The algorithm was used for networks of up to 2,000 nodes on a typical computer workstation, and certain data structures that save storage and solve the assignment problem for even a 5,000-node network are discussed.

As is well known traffic assignment is the process of finding the flow pattern in a given network with a given travel demand between the origin-destination (O-D) pairs. Equilibrium assignment finds flow patterns under user equilibrium, when no driver can unilaterally change routes to achieve better travel times. Optimal assignment determines the flow patterns such that the total travel time cost in the network is minimum, usually under external control. Assignment has long been an essential step in the transportation planning process. See Sheffi (1) for detailed discussions on traffic assignment.

Real-time traffic assignment could be a part of potential advanced traveler information systems or advanced traffic management system (ATMS) strategies [see Kaysi and Ben-Akiva (2) for a discussion of such strategies]. The applications of network assignment in such real-time contexts could be in the on-line estimation of O-D demand matrices or in an on-line dynamic assignment framework. The conventional approach, used in planning applications, is to solve the assignment problem with the Frank-Wolfe algorithm (F-W), also known as the convex-combinations algorithm (1). However real-time applications typically require path-based solutions [see Mahmassani and Peeta (3)], which are not available with the link-flow-based F-W. Faster convergence is also a very desirable feature for a real-time algorithm.

In this paper we report our investigation of the Goldstein-Levitin-Poljak gradient projection algorithm formulated by Bertsekas (4). This algorithm falls under the set of algorithms called *path-enumeration algorithms*, which have traditionally been discarded by transportation researchers as too memory intensive and slow for large networks. In light of the orders of magnitude improvement in the availability of computer memory in recent years, we believe that

¹ R. Jayakrishnan and S. Rajadhyaksha, Department of Civil and Environmental Engineering, University of California, Irvine, Irvine, Calif. 92664. W.K. Tsai, Department of Electrical and Computer Engineering, University of California, Irvine, Irvine, Calif. 92664. J.N. Prashker, Technion-Israel Institute of Technology, Haifa, Israel.

such algorithms deserve a fresh look. In this paper we describe our implementation of the algorithm and the extremely encouraging results. We discuss the assignment formulation for the sake of completeness in the next section, and follow it by a literature review and further qualitative discussions of the algorithms. We then proceed to discuss the computer implementation issues. We conclude with results on the comparative performances of the algorithms and pointers for future research.

STATIC USER EQUILIBRIUM ASSIGNMENT PROBLEM

As is well known the static assignment user equilibrium problem is stated as

$$\min Z = \sum_a \int_0^{x_a} t_a(\omega) d\omega \quad (1)$$

subject to the demand and nonnegativity constraints given by

$$\sum_k f_k^{rs} = q_{rs} \forall r, s, k \in K_{rs} \quad (2)$$

where

- x_a = flow on link a (sum of the flows on the paths sharing link a),
- $t_a(\omega)$ = cost (travel time) on link a for a flow of ω ,
- f_k^{rs} = flow on path k connecting origin r and destination s ,
- q_{rs} = total traffic demand between r and s , and
- K_{rs} = set of paths with positive flow between r and s .

The above problem or variations of the same problem have appeared in some of the recently proposed dynamic assignment algorithms with time-varying demands such as the bilevel algorithm of Janson (5) and the instantaneous dynamic assignment algorithm of Ran et al. (6). Note also that a system optimal assignment problem reduces to a user equilibrium problem with transformed (marginal) cost functions (1), and hence algorithms developed for user equilibrium assignment are applicable to the system optimal assignment as well.

REVIEW OF RELEVANT LITERATURE

Extensive work on network optimization approaches has been done to address the traffic equilibrium assignment problem. A detailed discussion of the conventional approaches to it is presented by Sheffi (1). A detailed study by Lupi (7) showed that F-W is superior to most other algorithms. Nagurney (8) compared F-W with the algorithm of Dafermos and Sparrow (9) and found the latter to be in general more efficient. There has been some research to improve the efficiency of F-W. Arezki and Vin Vliet (10) presented an analytic implementation of the PARTAN technique as applied to F-W and presented results indicating improvements over the original algorithm. LeBlanc et al. (11) and Florian et al. (12) showed how the PARTAN method

could be applied to the traffic network equilibrium assignment problem and showed improved convergence in real networks. Weintraub et al. (13) investigated a method of improving the convergence of F-W by making modifications on the step size. One of the most recent improvements was by Larsson and Patriksson (14) who employed simplicial decomposition approaches to the original F-W.

Algorithms for assignment based on Benders decomposition have also been developed by Florian (15) and Barton et al. (16). The projection-based algorithms that have been developed in the past include those by Pang and Chan (17) and Dafermos (18). There has, however, not been much drawn from the advances made in the parallel field of optimal flow assignment in computer communication networks. The gradient projection algorithm popularized by Bertsekas and Gallager (19) is one such algorithm that we investigate in the present study. In computer communication the networks are usually smaller than the large urban networks in which traffic assignments are carried out for planning purposes, and this may have been why transportation researchers have not paid enough attention to the research in that field.

SELECTION OF ALGORITHMS: HISTORICAL PERSPECTIVE

The choice of an appropriate algorithm for the traffic equilibrium assignment problem is guided by several criteria for selection, depending on the specific needs of the application, with the overriding criteria often being the memory requirements of the algorithm and its speed of convergence. These concerns become increasingly critical as the network size increases. We provide the following discussion to reveal the motivations behind the research.

The conventional choice for the traffic assignment problem so far has been F-W. This choice has been guided largely by the memory requirements criterion. Since F-W at any one iteration deals with only a single path between each O-D pair, its storage requirements are well within the capabilities of most ordinary computers. However it has the drawback that typically the convergence becomes very slow as it approaches the optimal solution. It shows a tendency to flip-flop as it gets close to the optimum. The reason is that the algorithm is driven more by the constraint corners and less by the actual descent direction of the objective function surface once it is close to the solution. This was not considered a serious problem in earlier applications of the traffic assignment because the problem was being addressed from a transportation planning viewpoint. Under this scenario assignment is used for forecasting purposes when the O-D demand data themselves are derived by using the extrapolation of current values or statistical models. This inherent inexactness in the process renders the exact estimation of link volumes unimportant, and so practitioners are often content to stop the algorithm after a few iterations when it reaches within 5 to 10 percent of the solution. The memory requirement criterion was also of importance. When F-W was introduced to the transportation field in the late 1970s computers were incapable of handling the larger memory requirements of path-enumeration algorithms. Recent advances in computing equipment have placed vastly increased computing power in smaller and smaller machines. Computer workstations with 16 megabytes of storage are only about as expensive as a personal computer with 128 kilobytes of storage in the mid-1980s and have more storage than the largest mainframe computers of the late 1970s. Given these possibilities it is important that we rethink our choice of traffic assignment algorithms and

take advantage of the technological edge provided by current and future improvements in computer hardware.

Another aspect of F-W is that it does not automatically find the intersection turning movements. This has traditionally been found by microcoding the intersections with specific turning links, which usually increases the numbers of nodes and links in the networks considerably. Path-flow solutions automatically provide such turning counts without any microcoding of the network, thus keeping the network sizes small. However if separate flow-cost functions are to be used for turning movements, microcoding may be necessary with path-based algorithms also.

The recent advent of the intelligent vehicle-highway system (IVHS) brings up the need for real-time traffic assignments with requirements different from those for planning applications. Such assignments may be part of dynamic assignment frameworks or real-time O-D demand estimation frameworks. Faster convergence becomes important, and path-based solutions may be necessary. Moreover increasing emphasis is placed on estimating fuel consumption and modeling air quality over specific routes. Solutions based on path flows provide speed profiles over the network paths that are conceivably useful (although still very approximate) for such applications. Link-flow solutions from F-W are much poorer in this regard.

GRADIENT PROJECTION ALGORITHM

The gradient projection algorithm (GP) is extensively used in computer communication networks, in which path-flow solutions are essential for optimal flow routing. However the networks in these applications are typically much smaller than urban traffic networks and the path-enumeration issues have not been serious concerns. Moreover the network structures are also somewhat different in these two applications. We adapted the basic Goldstein-Levitin-Poljak GP formulated by Bertsekas (4) to the traffic assignment problem and concentrate here on the practical convergence and computer implementation issues.

In contrast to F-W, which finds auxiliary solutions that are at the corner points of the linear constraint space, GP makes successive moves in the direction of the minimum of a Newton approximation of a transformed objective function. The objective function includes the demand constraints also, and thus the feasible space for gradient projection is defined only by the nonnegativity constraints, as opposed to both nonnegativity and demand constraints in the case of F-W. Should the move to the minimum in the negative gradient direction result in an infeasible solution point, a projection is made to the constraint boundaries. As a result of the redefinition of the problem, infeasibility occurs only when a variable violates the nonnegativity constraint, and thus the projection is easily accomplished by making that variable zero. We describe this in detail below.

The formulation of the algorithm focuses on the traffic demand constraints.

$$\sum_{k \in K_{rs}} f_k = q_{rs}$$

where K_{rs} is the set of paths (with positive flow) between origin r and destination s .

If we express the shortest-path flows $f_{k_{rs}}^-$ in terms of other path flows

$$f_{\bar{k}_{rs}} = q_{rs} - \sum_{\substack{k \in K_{rs} \\ k \neq \bar{k}_{rs}}} f_k \quad (4)$$

the standard optimization problem (equations 1 through 3) can be restated as

$$\min \bar{Z}(\bar{f}) \quad (5)$$

subject to

$$f_k \geq 0 \quad \forall f_k \in \bar{f} \quad (6)$$

where \bar{Z} is the new objective function, and \bar{f} is the set of non-shortest-path flows between all of the O-D pairs.

For each O-D pair while at any feasible (nonoptimal solution) a better solution can be found by moving in the negative gradient direction. This gradient is calculated with respect to the flows on the non-shortest paths (which are the only independent variables now), and a move size is found by using the second derivatives with respect to these path-flow variables. Once the flows on these non-shortest paths are updated the flow on the shortest path is appropriately updated so that the demand constraint is satisfied.

The gradient of the objective function written in terms of the non-shortest-path variables can be found using

$$\frac{\partial \bar{Z}}{\partial f_k} = \frac{\partial Z}{\partial f_k} - \frac{\partial Z}{\partial f_{\bar{k}_{rs}}} \quad \text{where } k \in K_{rs} \text{ and } k \neq \bar{k}_{rs} \quad (7)$$

which results from the definition of \bar{Z} . Thus each component of the gradient vector is the difference between the first derivative lengths of a path and the corresponding shortest path (14). In the case of equilibrium assignment the objective function is in terms of integrals and the first derivative lengths are simply the path costs at that flow solution.

A small increase in the flow on a path k results in an equal decrease in the flow on the corresponding shortest path. This results in no change in the flow on the common part of the two paths. Thus the second derivative is simply the sum of the second derivative lengths of the links on either path k or path \bar{k}_{rs} , but not both. A small increase in the flow on path k causes an equal

decrease in the flow on the shortest-path \bar{k}_{rs} . The flows on the common links on these paths do not change. The increase in flow on the other links on path k causes positive second derivatives. The decrease in flow on the other links on \bar{k}_{rs} , also causes positive second derivatives as it increases the negative first derivatives. Once the second derivatives of \bar{Z} with respect to each path flow are calculated, we assume a diagonal Hessian matrix, and the inverse of each second derivative gives an approximate quasi-Newton step size for updating each path flow.

For the remainder of the paper when we refer to the *first derivative lengths* we mean the first derivatives of the objective function, which is composed of link costs at specific path flows [i.e., $t_a(x_a)$]. Similarly, *second derivative length* refers to the second derivative of the objective function and is composed of first derivatives of link costs (i.e., $\partial_d/\partial x$ at $x = x_a$).

On the basis of the above discussion the gradient projection algorithm can be formalized as follows:

Step 0 – Initialization: Set t_a equal to $t_a(0)$, $\forall a$ and perform all-or-nothing assignments. This yields path flows f_l^{rs} , $\forall r, s$ and link flows x_a^l , $\forall a$. Set iteration counter n equal to 1. Initialize the path set K_{rs} with the shortest path for each O-D pair rs .

Step 1 – Update: Set t_a equal to $t_a(x_a^n)$, $\forall a$. Update the first derivative lengths d_k^n (i.e., path costs at current flow) of all of the paths in K_{rs} $\forall r, s$.

Step 2 – Direction finding: Find the shortest-path \bar{k}_{rs}^n from each origin r to each destination s on the basis of $[t_a^n]$. If different from all the paths in the existing path set K_{rs} (no need for path comparison here; just compare d_k^n), add it to K_{rs} and record $d_{\bar{k}_{rs}^n}^n$. If not tag the shortest among the paths in K_{rs} in $d_{\bar{k}_{rs}^n}^n$.

Step 3 – Move: Set the new path flows.

$$f_k^{n+1} = \max \left[0, f_k^n - \frac{\alpha^n}{s_k^n} (d_{\bar{k}_{rs}^n}^n - d_{\bar{k}_{rs}^n}^n) \right] \forall r, s, k \in K_{rs}, k \neq \bar{k}_{rs}^n$$

where

$$s_k^n = \sum_a \frac{\partial_a^n}{\partial x_a^n} \forall k \in K_{rs}$$

a denotes links that are on either k or \bar{k}_{rs}^n but not on both, and α^n is a scalar step-size modifier (say, $\alpha^n = 1$).

Also,

$$f_{\bar{k}_{rs}^n}^{n+1} = q_{rs} - \sum_k f_k^{n+1} \quad \forall k \in K_{rs} k \neq \bar{k}_{rs}^n$$

Assign the flows on the trees and find the link flows x_a^{n+1} .

Step 4 – Convergence test: If the convergence criterion is met, stop, or set n equal to $n + 1$ and go to Step 1.

It is better to keep α^n a constant (i.e., $\alpha^n = \alpha, \forall n$). It can be shown that given any starting set of path flows there exists an $\bar{\alpha}$ such that if $\alpha \in (0, \bar{\alpha})$ the sequence generated by this algorithm converges to the optimum (1), provided that the link-cost functions are convex. Our experience shows that α equal to 1 achieves a very good convergence rate, and all of the results in this paper use this value of α . The solutions reached are unique in terms of the link flows, but the path-flow solution, although it is optimal, is not necessarily unique [see Sheffi (1) for a discussion on why the path-flow solutions need not be unique].

A qualitative graphical comparison of GP and F-W is shown in Figure 1. In this case F-W moves in directions that are almost orthogonal to the descent direction, once it is close to the optimum, because the moves are toward constraint corners to avoid infeasibility. GP still moves in the descent direction when it is closer to the optimum. Note that this is just an example and is provided only to illustrate the qualitative reason behind the faster convergence of GP. The actual nature of the objective functions and the constraints in the network assignment context are quite different.

GP distributes flows from existing paths to the shortest path during every iteration, with different fractions of flow being taken out of the alternative paths between an O-D pair. A careful look at F-W shows that it also implicitly redistributes flows from alternative paths. However the fractions taken out from the paths are all same, and the path-flow solutions are never kept track of.

COMPUTATIONAL STORAGE CONSIDERATIONS

GP is a path-enumeration algorithm, and the paths need to be carefully stored to prevent memory problems. This is an issue that has rarely been addressed in the computer communication applications, but we address this here because we deal with traffic networks with large sizes. In GP one shortest-path tree is built during each iteration from each origin. The paths between an O-D pair are all generated during different iterations, and each path is part of one of the shortest-path trees built from the corresponding origin node. It is important not to store the paths as node lists but rather as predecessor trees (note that each iteration produces the shortest paths, which are invariably trees). This avoids double storage of the common portions of different paths found from each origin in each iteration. As we store one predecessor node number for each node, each tree in a network of N nodes requires N storage locations. This results in $N_o \cdot N$ storage locations in each iteration, where N_o is the number of origins. Thus the main memory requirement of the algorithm is of the order of $N_o \cdot N \cdot N_i$, where N_i is the number of iterations. We do not see the

kind of combinatorial explosion that is expected to occur with path-enumeration algorithms. The fact that the paths in each iteration are part of trees is thus a very handy feature of GP.

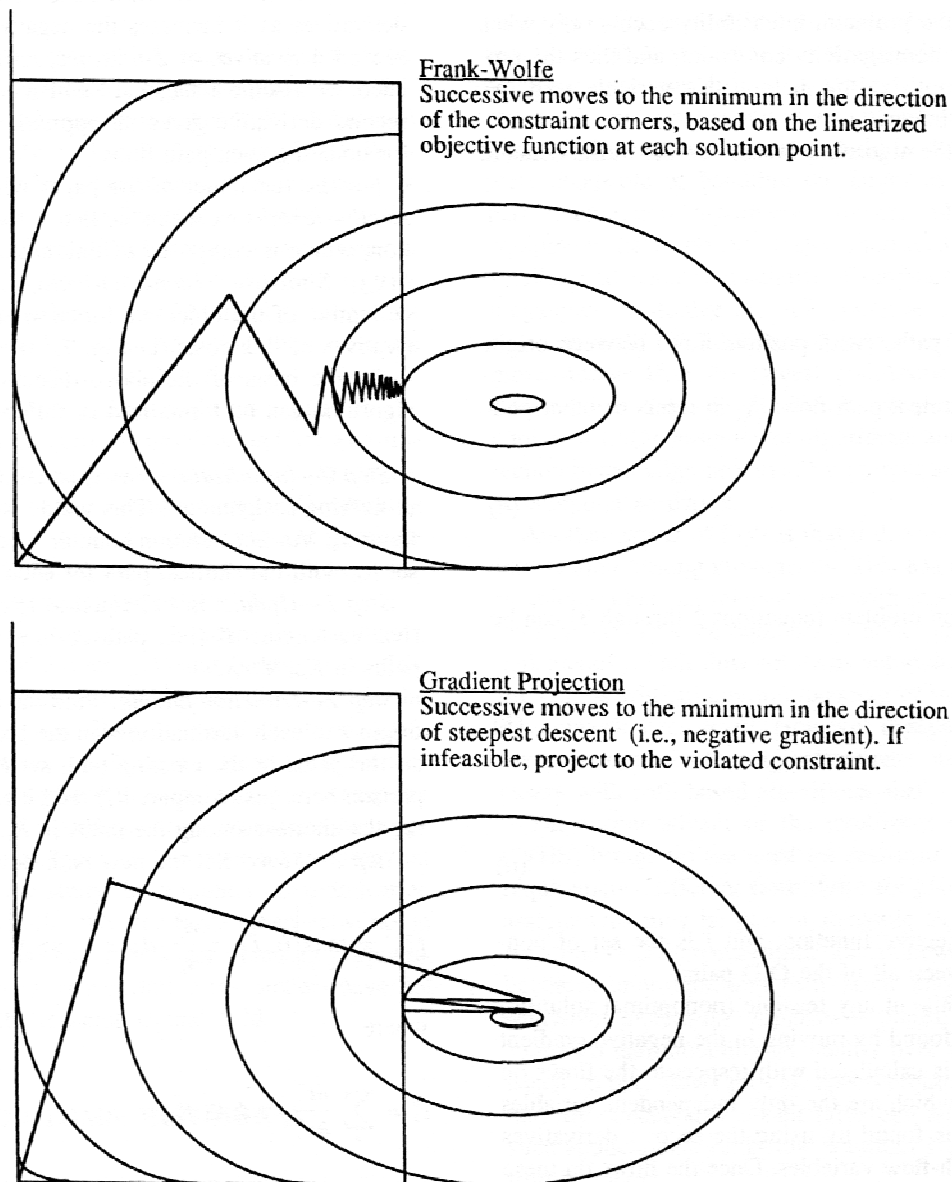


FIGURE 1 Comparison of GP and F-W.

However for F-W the memory requirements are fixed by network size and are not affected by the number of iterations until convergence. Typically its requirements are of the order of $N_o \cdot N$ because it stores only one tree (the shortest paths for the all-or-nothing assignment) in every iteration. Thus GP does have a storage disadvantage compared with F-W, but it is able to provide a richer solution for precisely the same reason because it gives us path flows, as opposed to the link flows provided in F-W.

The memory requirement is hardly a significant concern on the basis of our experience. With a simple predecessor array data structure we were able to run networks of up to 1,200 nodes with 22,300 O-D pairs on a SUN workstation. If only about 10 iterations of GP are attempted (which itself generally finds better solutions than 100 iterations of F-W, as our results shown in this paper indicate), we can run networks of more than 2,000 nodes. We briefly describe a new data structure that allows us to run networks of up to 5,000 nodes and 560,000 O-D pairs to 32 iterations. The purpose for attempting to run such networks is to show that the storage problem that kept researchers away from applying path-enumeration algorithms to larger networks is really not a problem anymore, at least in the case of GP.

The following describes an efficient data structure for larger network problems. The shortest-path trees built in successive iterations often have several identical branches. When these are stored as separate trees it results in a great amount of duplication of storage because we find that several nodes have the same predecessor arcs in successive trees. Rather than storing the predecessor arcs for all of the nodes in every iteration we store a shorter list of nodes for which the predecessor arcs change in an iteration (change being defined from the predecessor in the "anchor" iteration, say, the first one). The information regarding the iteration numbers in which the predecessor of each node changes is stored in terms of the bits of a number. A 1 in bit location i of this number for a node means that its predecessor changed in iteration i , and a 0 would indicate that no change occurred in iteration i . In a computer with 32 bit numbers, this lets us store the information with just N numbers. To find the predecessor for this node during, say, iteration j , we need to find the bit location of the last 1. This can be efficiently done at the hardware or software level and yields the appropriate iteration number, i . We go to the short list of changed predecessors corresponding to iteration i to find the predecessor for the node of concern. This approach will not achieve good computation time results unless it is carefully implemented, and we leave out the complicated implementation details in this paper. With this data structure we were able to reduce the storage requirement for the trees from the $N_o \cdot N \cdot N_i$ above, to about $N_o \cdot N \cdot C$, where C is about 5 to 10 for up to even 100 iterations (i.e., $N = 100$). This is because the predecessors of each node change only fewer than 10 times during 100 iterations of GP on the basis of our assignment runs on realistic traffic networks.

COMPUTATION TIME CONSIDERATIONS

A careful implementation is absolutely essential for GP to perform well. Because we found that the algorithm converges generally about 10 times faster than F-W in terms of the number of iterations, our intention was to ensure that GP achieves similar speed in computation time also. Although F-W requires no other operations of computational intensity comparable to that for the shortest-path determination during each iteration, GP has other procedures during its iterations that can be more time intensive than the shortest-path determination for larger networks. Our studies (as discussed in the next section) show that GP converges 10 times faster than F-W for a 100-node network but only 60 percent faster for a 900-node network, although the number of iterations needed is still less than one-tenth. Because we found that almost all of the time in F-W is spent on the shortest-path routine for larger networks, this means that there are routines in GP whose computational intensities increase faster than that for the shortest-path routine as the network size increases. We identify three such key operations, and these must be carefully implemented: (a) assigning the flow on each path to the links along its length to find the total link

flows, (b) finding the first derivative lengths for all of the paths between each O-D pair, and (c) finding the second derivative lengths for each pair. We have been successful in developing efficient schemes for the first two but have not been able to tackle the third problem without modifying the algorithm itself. The results that we provide in this paper are based on a program that includes only the techniques for the first derivative lengths. However we discuss all three aspects here to show the potential of the algorithm to show even better results than we have. Provided if our suggestions are implemented. Our early results with all three of the following procedures are indeed very encouraging.

Implementation of Flow Assignment Procedure

Implementation of the flow assignment procedure refers to assigning the path flow to all of the links on each path after the path flows are updated during each iteration. There are $N_o \cdot N_d$ O-D pairs in a network (where N , is typically 10 to 20 percent of N), and the expected number of active paths between an O-D pair, N_p , is typically about 5 to 10 at convergence. Each path in a traffic network has roughly $O(N^{1/2})$ links on them. Thus this operation could be of $O(N_o \cdot N_d \cdot N^{1/2} \cdot N_p)$ effort in each iteration, if each path is considered independently (i.e., the link-level operations are repeated for paths that share common portions). in contrast to this an efficient implementation (say, using a heap) of a routine to find the shortest paths from all origins to all nodes results in $O(N_o \cdot N \log N)$ or better computational intensity. The flow assignment step becomes much more time-consuming for larger networks. We developed an efficient tree-traversal procedure to assign the flows instead of doing it path by path. The procedure starts from a leaf node of the tree and goes up assigning the flow on the links until it reaches a node where there is another branch with no flow assigned. Once the flows are added on all of the branches at a node we find the total flow that should be assigned on the predecessor link of the node and move up. This procedure goes only once over each arc in the tree, and hence it is an $O(N)$ operation for each tree (a maximum of four additions per arc in a typical traffic network). This results in only $O(N_o \cdot N)$ operations for all origins, all destinations, and all paths in each iteration. We leave out further details of this procedure for brevity. It suffices to say that this assigns flows of multiple paths sharing common portions without repeated calculations at the links. This is a significant improvement because the computations now do not depend on the number of paths or the number of nodes on the paths. The computational intensity drops to below that for the shortest-path determination with this method.

Finding First Derivative Path Lengths

In finding the first derivative path lengths the link costs are added up on different paths. Similar to the above, we need to avoid path-based computations of the $O(N_o \cdot N_d \cdot N^{1/2} \cdot N_p)$ order in this case also. Here we again perform a tree traversal procedure. We go up from any node in the tree until we reach the root (origin) node and add the link costs on the links to find the first derivative length to that node. Then we go up from that node once more (second pass), subtracting one link cost at a time to find the costs to each node along the way. The two passes are needed only because we cannot move down the tree when the tree is stored with predecessor representation. A threaded-tree storage will let us do a one-pass traversal, but additional overheads may be involved. Another option is to keep the tree traversal order right after each tree is built, but this

requires as many storage locations as the tree itself and doubles the storage requirements. Then we move to any node not yet considered and repeat the procedure, but this time starting the second pass after reaching the origin node or a node with an already-computed path length. Because each node is reached strictly twice, this results in only $O(N_o \cdot N)$ operations in every iteration, which is much faster than the shortest-path determination in larger networks.

Second Derivative Length Calculations

Second derivative length calculations require the addition of second derivative lengths of links not common between each path and the corresponding shortest path. If this is done path by path, adding the second derivatives on each link with the shortest path, this also results in $O(N_o \cdot N_o \cdot N^{1/2} \cdot N_p)$ computations. We have so far not been able to find an $O(N_o \cdot N)$ technique for this without changing GP itself to some extent. The difficulty arises because the path under consideration is on another tree that is different from the tree of which the current shortest path is a part. It is possible that we can improve the situation only by changing the algorithm substantially. We suggest the use of a line search rather than the second derivatives to find the step size in the negative gradient direction. An auxiliary path-flow solution can be easily found in the negative gradient direction, and then an unconstrained line search can be used to determine the step size to reach the minimum in this direction. This line search can be performed fast in the link-flow domain (using the link flows at the current and auxiliary path-flow solutions), and on the basis of the optimal step size a path-flow update is performed. The flow update would be based on path flows. Our early experience with this technique has been encouraging.

ASSIGNMENT RESULTS

The assignment studies compare the performance of GP with that of F-W. To make conclusions on the comparative performances of the algorithms it is necessary that they be tested under sufficiently diverse networks. We studied the algorithms on grid networks of different sizes generated by using a random network generator program that we developed as well as on the network of major arterials and freeways in Anaheim, California.

The test networks are grids only in terms of the connectivities of the links, with the link lengths being determined randomly. There are two links each way between the nodes. The link lengths are randomly picked from a uniform distribution of between 500 and 5,000 ft. The free-flow speeds on the links are randomly picked from a uniform distribution of between 22 and 40 mph. The capacities of the links are based on the number of lanes (one, two, or three), with each lane having a capacity of 1,800 vehicles per hr. Certain nodes from the network are randomly picked as O-D centroids. This is done on the basis of a set of rules that attempts to create a network representative of real-world traffic networks. First approximately 12.5 percent (one-eighth) of the total nodes in the network are picked to be centroid nodes, which is about the fraction of zone-centroid nodes in typical assignment applications. There are at least three links between any two centroid nodes to ensure that they are not too close to each other. Once the centroids are set up the O-D flow matrix is generated. Each centroid generates demand at a prespecified rate (9,600 vehicles per hr was used in our studies), and the generated traffic is distributed to other nodes on the basis of the inverse squared distances to develop the O-D matrix.

The Anaheim network has 416 nodes (of which 38 are O-D centroids), 914 arcs, and 1,406 O-D pairs. A static O-D demand matrix was estimated by using the COMEST program on the basis of some link counts in the network. The demand data refer to the evening peak period in the network, which has moderately high levels of congestion. No microcoding of the intersections was attempted for this network.

The assignments were carried out by using a Bureau of Public Roads link-cost function, $t = t_o[1 + 0.15 (x/c)^4]$, where t is the link travel time cost, t_o is the free-flow cost, x is the flow, and c is the link capacity. Both GP and F-W included identical shortest-path routines, which is based on a binary heap data structure. The line search routine for F-W uses an efficient Bolzano search [see Sheffi (1)]. The programs were implemented in FORTRAN-77 on a Sun SPARC-11 workstation with 64 megabytes of storage. The flows and costs were floating point variables.

Table 1 shows the results from assignments on networks of various sizes. For all of the network sizes, Table 1 shows the number of iterations required by F-W to find the objective function value that GP finds in 2, 4, 6, 8, and 10 iterations as well as the corresponding computation times. F-W requires 30 to 160 iterations to reach the solutions found by GP in just 6 iterations. For all of the networks we found that GP converges between 10 and 15 iterations to solutions that F-W takes between 300 and 2,000 iterations to reach. There is at least an order of magnitude improvement in terms of the number of iterations.

The computation times are also improved similar to the reduction in iterations for smaller networks. This is expected because the main computational step is the shortest-path determination for both GP and F-W in small networks. However the computation times with GP are about 40 percent of those with F-W for 10 iterations of GP in a 1,000-node network. This shows that procedures other than the shortest-path determination use up significant time in GP for larger networks, as explained in the previous section. It should be stressed that these assignments were carried out with an implementation of GP that does not yet include most of the procedures that we explained before. Thus even though a 60 percent improvement is significant, the computation times for GP can be reduced even further than those shown in Table 1, especially for larger networks.

Figures 2 to 5 show the results of the assignments on the network of Anaheim for various demand levels. The demands generated by the O-D matrix estimated from actual link counts are denoted as a demand level of 1.0. For other demand levels the cells in the trip table were all multiplied by appropriate fractions. These assignments were carried out to examine the effect of the demand level on the relative performances of GP and F-W. Again we see that for all cases F-W takes more than 10 to 15 iterations to reach the solutions found by GP in 1 to 3 iterations. There does not appear to be a significant change in the performance of GP in comparison with that of F-W as the demand level increases. Both algorithms require more iterations to converge for higher demand levels, but GP still shows 5 to 10 times faster convergence.

TABLE I Comparative Performances of GP and F-W						
Square Grid Networks	Gradient Projection			Frank-Wolfe		
	Iterations	Obj. function	Time (sec)	Iterations	Obj. function	Time (sec)
36 nodes	2	5385.2	0.08	4	5376.7	0.08
120 arcs	4	4890.2	0.14	27	4889.2	0.51
12 OD pairs	6	4794.5	0.17	54	4793.9	1.02
	8	4747.9	0.22	123	4747.8	2.32
	10	4728.8	0.28	390	4728.8	7.35
100 nodes	2	13076.9	0.38	5	13035.8	0.44
360 arcs	4	12562.6	0.76	30	12561.9	2.64
132 OD pairs	6	12526.0	1.14	168	12526.0	14.81
	8	12522.1	1.52	618	12522.1	68.58
	to	12521.4	1.90	1282	12521.4	198.32
225 nodes	2	33872.1	1.87	7	33757.4	2.73
840 arcs	4	32979@2	4.13	36	32977.7	14.03
756 ODs	6	32912.4	6.26	176	32912.4	68.58
	8	32904.8	8.44	509	32904.8	198.32
	10	32902.3	10.52	1611	32902.3	627.20
400 nodes	2	70849.9	5.78	6	70835.1	6.90
1520 arcs	4	68797.2	13.86	21	68774.8	24.14
2450 ODs	6	68408.2	22.67	65	68407.5	74.71
	8	68343.9	31.33	185	68343.9	212.63
	10	68326.8	38.97	537	68326.8	617.19
625 nodes	2	124874	14.51	7	124225	19.03
2400 arcs	4	120151	38.56	23	120144	62.52
6006 ODs	6	119322	64.04	48	119315	130.49
	8	119100	88.77	88	119100	239.22
	10	119028	112.39	156	119028	424.08
900 nodes	2	206379	30.90	7	206097	39.56
3480 arcs	4	198483	86.53	26	198436	146.93
12432 ODs	6	197974	148.76	31	197929	175.19
	8	196668	207.98	80	196665	452.10

	10	196508	271.74	123	196508	695.10
--	----	--------	--------	-----	--------	--------

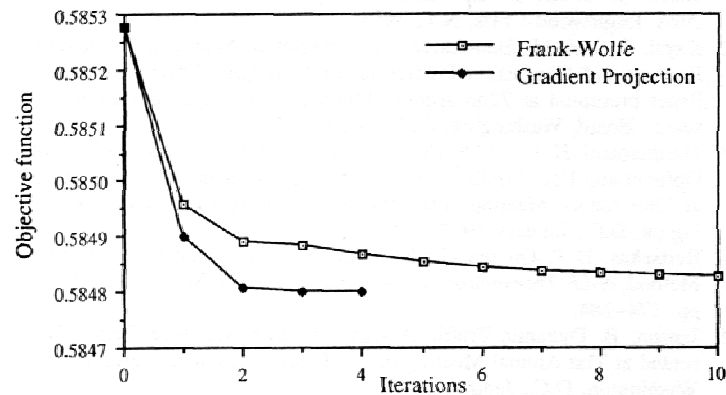


FIGURE 2 Comparison of GP and F-W on Anaheim, California, network (volume level = 0.5).

We did not compare the PARTAN version of F-W with GP. However, published results on PARTAN (10,11) indicate that this typically is about twice as fast as ordinary F-W, in the number of iterations, in finding solutions. On the basis of the improvements that we found with GP we decided that the comparison of the PARTAN version of F-W with GP was not warranted at this time. Moreover PARTAN is still not commonly used for assignments by practitioners. We do intend, however, to carry out these comparisons in the future.

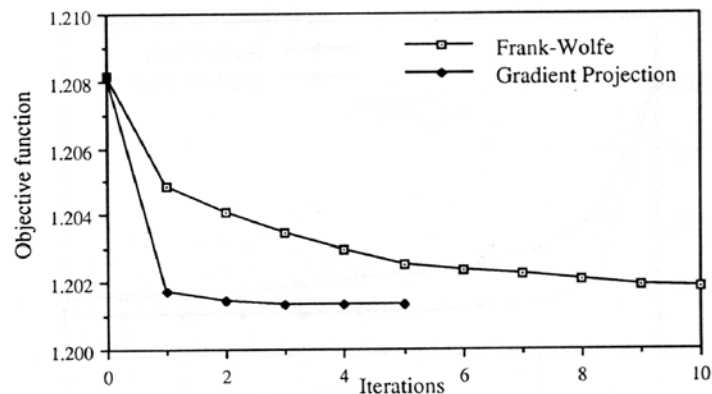


FIGURE 3 Comparison of GP and F-W on Anaheim, California, network (volume level = 1.0).

CONCLUSIONS

In this paper we provided a detailed discussion and supportive results to show that path-enumeration algorithms such as gradient projection deserve a fresh look for applications in traffic assignment. There were two main motivations behind the research: (a) the tremendous improvement in recent years of the availability of computer memory, and (b) the need for fast

assignment algorithms for certain possible IVHS strategies for optimal routing and guidance on the basis of dynamic assignment frameworks, real-time trip table estimation, and so on. We show that path-based algorithms can be applied to networks of thousands of nodes. We provide data structures that can be used to handle path-based storage problems, and we suggest techniques for achieving the fast completion of path-based procedures in the algorithm. These techniques are also applicable to other path-based algorithms. Our implementation of GP converges in an order of magnitude fewer iterations than conventional F-W and can be made to show similar computation time speedup if implemented carefully.

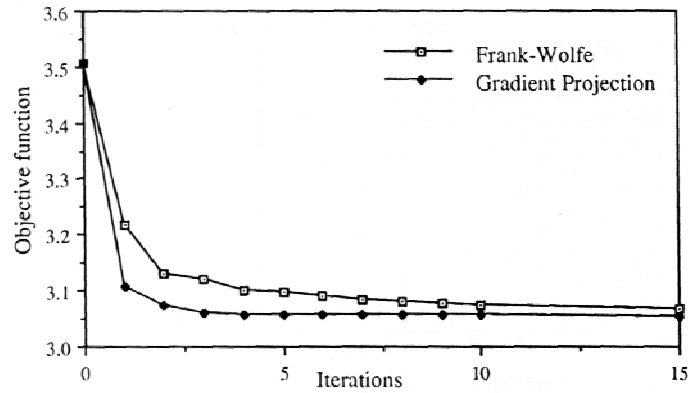


FIGURE 4 Comparison of GP and F-W on Anaheim, California, network (volume level = 1.5).

There are advantages to the path-based solutions generated by GP. Such solutions can be used directly in path-based routing frameworks. Another advantage is the direct determination of node turning counts without microcoding the intersections and increasing the network size. In addition we can find the link-to-link flow variation on each path. This may provide some opportunities for finding approximate estimates of fuel consumption, environmental impacts, and so on, for selected paths or O-D pairs.

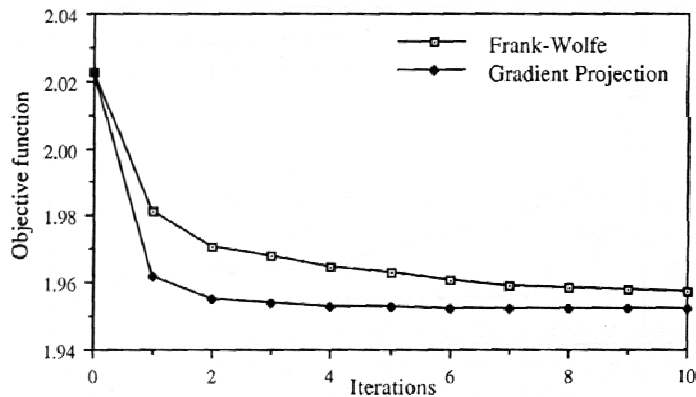


FIGURE 5 Comparison of GP and F-W on Anaheim, California, network (volume level = 2.0).

Several aspects of the algorithm require further study. One important aspect is the convergence rates under different link-cost functions. Although we have carried out some research in this area and have found the results to be reasonably robust, our research has by no means been exhaustive. Application to other related problems such as dynamic assignment and variable demand assignment would provide more insights on the algorithm's performance. Research is also under way at the University of California, Irvine, on developing gradient projection with hierarchical decomposition schemes for traffic network assignment.

ACKNOWLEDGMENTS

We wish to thank Wilfred Recker, Institute of Transportation Studies at the University of California, Irvine, for his constant encouragement. The research was supported by the California Department of Transportation under the Anaheim ATMS test bed research program, and we acknowledge the generous support.

REFERENCES

1. Sheffi, Y. *Urban Transportation Networks*. Prentice-Hall, Incorporated, Englewood Cliffs, N.J., 1985.
2. Kaysi, I., and M. Ben-Akiva. An Integrated Approach to Vehicle Routing and Congestion Prediction for Real-Time Driver Guidance. Paper presented at 72nd Annual Meeting of the Transportation Research Board, Washington, D.C., January 1993.
3. Mahmassani, H. S., and S. Peeta. Network Performance under System Optimal and User Equilibrium Dynamic Assignment. Paper presented at 72nd Annual Meeting of the Transportation Research Board, Washington, D.C., January 1993.
4. Bertsekas, D. P. On the Goldstein-Levin-Poljak Gradient Projection Method. *IEEE Transactions on Automatic Control*, Vol. AC-21, 1976, pp-174-184.
5. Janson, B. Dynamic Traffic Assignment with Schedule Delay. Presented at 71st Annual Meeting of the Transportation Research Board, Washington, D.C., January 1992.
6. Ran, B., D. E. Boyce, and L. J. Leblanc. A New Class of Instantaneous Dynamic Assignment Models. *Operations Research*, Vol. 41, No. 1, 1993.
7. Lupi, M. Convergence of the Frank-Wolfe Algorithm in Transportation Networks. *Civil Engineering Systems*, Vol. 19, 1985, pp. 7-15.
8. Nagurney, A. B. Comparative Tests of Multimodal Traffic Equilibrium *Methods*. *Transportation Research B*, Vol. 18B, No. 6, 1984, pp. 469-485.
9. Dafermos, S. C., and F. T. Sparrow. The Traffic Assignment Problem for a General Network. *Journal of Research of the National Bureau of Standards-B*. Vol. 73B, No. 2, 1969, pp. 117-119.

10. Arezki, Y, and D. Van Vliet. A Full Analytical Implementation of the PARTAN/Frank-Wolfe Algorithm for Equilibrium Assignment. *Transportation Science*, Vol. 24, 1990, pp. 58-62.
11. LeBlanc, L. J., R. V Helgason, and D. E. Boyce. Improved Efficiency of the Frank-Wolfe Algorithm for Convex Network Problems. *Transportation Science*, Vol. 19, 1985, pp. 445-462.
12. Florian, M., J. Guélat, and H. Spiess. An Efficient Implementation of the PARTAN Variant of the Linear Approximation Method for the Network Equilibrium Problem. *Networks*, Vol. 17, 1987, pp. 319-339.
13. Weintraub, A., C. Ortiz, and J. Gonzales. Accelerating Convergence of the Frank-Wolfe Algorithm. *Transportation Research B*, Vol. 19B, 1985, pp. 113-122.
14. Larsson, T., and M. Patriksson. Simplicial Decomposition with Disaggregated Representation for the Traffic Assignment Problem. *Transportation Science*, Vol. 26, 1992, pp. 4-17.
15. Florian, M. A Traffic Equilibrium Model of Travel by Car and Public Transit Modes. *Transportation Science*, Vol. 11, 1977, pp. 166-179.
16. Barton, R. B., D. W. Hearn, and S. Lawphongpanich. The Equivalence of Transfer and Generalized Benders Decomposition Methods for Traffic Assignment. *Transportation Research B*, Vol. 23B, 1989, pp. 61-73.
17. Pang, J. S., and D. Chan. Iterative Methods for Variational and Complementarity Problems. *Mathematical Programming*, Vol. 24, 1982, pp.284-313.
18. Dafermos, S. C. A Variable Inner-Product Projection Algorithm for Variational Inequalities with Application to the Traffic Equilibrium Problem. Working paper. Lefschetz Center for Dynamical Systems, Brown University, Providence, R.I., 1983.
19. Bertsekas, D. P., and R. G. Gallager. *Data Networks*. Prentice-Hall, Incorporated, Englewood Cliffs, N.J., 1987.

Publication of this paper sponsored by Committee on Transportation Supply Analysis.

Cost Versus Time Equilibrium over a Network

FABIEN LEURENT¹

Most traffic assignment models assume that the generalized cost experienced by a traveler making a given trip on a network results from a combination of time and monetary expenses that is the same for everybody. To represent disaggregate trade-offs between time and monetary expenses, a model that differentiates travelers by means of an attribute value of time, was designed. It was assumed that this attribute is continuously distributed across the population of trip makers. After giving the economic foundation of the cost-versus-time model with continuous values of time, variable demand, and congestion effects on travel times, it is mathematically characterized as a solution of a convex minimization program. Then existence and uniqueness results as well as a convenient algorithm that avoids path storage are provided. Finally a small numerical example that demonstrates the relevance of considering continuously distributed values of time when evaluating toll highway projects is presented.

Traffic assignment is an important part of the transportation planning process enabling one to simulate the trips made by people faced with a given transportation network. The models used to design new network facilities or to test new policies generally assume that all people experience the same generalized time on a given route, making a uniform trade-off between cost and time expenses.

For the evaluation of toll road projects that have mushroomed in France's largest towns, the differentiation of people according to their value of time (VOT, an attribute used to convert time into money) has proved an important advancement. Explicit modeling of the trade-offs between cost and time provides a more realistic way of simulating the users' responses to toll charges.

A first approach is to use a stochastic assignment model by having the random part of the utility account for the dispersion of trip-makers' VOTS. Both the logit model (1) and the probit model (2) can be adapted to that purpose. However if it is recognized that the VOT and its dispersion have a sound behavioral basis then a modeler should try to account for them analytically.

A second line of attack also consists of differentiating several classes of motorists, each one with a given VOT. The theoretical framework for the multiple user classes model has been worked out b), Dafermos (3) in the deterministic case and Daganzo (4) in the stochastic case. An implementation is available in the SATURN package (5).

In France most interurban mode choice models are related to the second methodology, with the only difference being that the VOT is assumed to be continuously distributed across the tripmakers (6-9). Such models are known as *cost versus time models* (*modèles prix-temps* in French). Before adapting those models to urban path choice, congestion effects on travel times should be considered. There have been some attempts (10,11) to develop equilibrium

¹ Département Economie et Sociologie des Transports, Institut National de Recherches sur les Transports et leur Sécurité, Avenue du Général Malleret-Joinville, 2, 94114 Arcueil, France.

assignment models able to compute a cost-versus-time equilibrium with travel times that depend on traffic flows. The theoretical background as well as the algorithms are heuristic.

To end the state-of-the-art review, a paper by Dial (12) should be mentioned. The paper presents a cost-versus-time model with a view to addressing both mode and route choices, but congestion is not taken into account.

The purpose of this paper is to introduce a cost-versus-time equilibrium model with variable demand, continuously distributed VOT, and flow-dependent travel time functions. This model can be used to study the potential traffic on urban toll roads and to assess middle- and long-run predictions owing to the variability of demand in the medium and long terms.

The remainder of the paper comprises four parts. First, the economic background of the cost-versus-time model is set. Second, the mathematical framework required to ensure the consistency of the model and to derive existence and uniqueness results is given. A convex programming characterization of a cost-versus-time equilibrium is provided. This section may be skipped by readers who are not interested in technical issues. Third, an algorithm to compute the cost-versus-time equilibrium is designed. It is convenient because it avoids path storage and enumeration. Finally a short example of an evaluation of a toll highway project is provided; it shows that an aggregate (single VOT) model gives results (specifically for the optimal toll and toll revenues) that are substantially different from those of the true, disaggregate cost-versus-time model.

ECONOMIC ISSUES

Modeling Disaggregate Cost-Versus-Time Trade-offs

If i is a trip-maker with VOT v_i and k is a path with travel time T_k and travel cost (price) P_k , the generalized travel time $G_k(i)$ experienced by the i th traveler on path k results from a combination of time and money expenses:

$$G_k(i) = T_k + \frac{P_k}{v_i} \quad (1)$$

A utility-maximizing trip-maker will travel on the path that exhibits the minimum generalized travel time to his or her own point of view.

If there are only two alternative paths, the first one cheaper but slower and the second one faster but more expensive, people with high VOTs would choose the second path, whereas people with low VOTs would be satisfied with the first one. Taking a French interurban mode choice example, the first path may be thought of as a train and the second path as an airplane.

The frontier VOT v^* between the two paths is such that it equalizes their generalized times:

$$T_1 + \frac{P_1}{v^*} = T_2 + \frac{P_2}{v^*} \quad (2)$$

hence

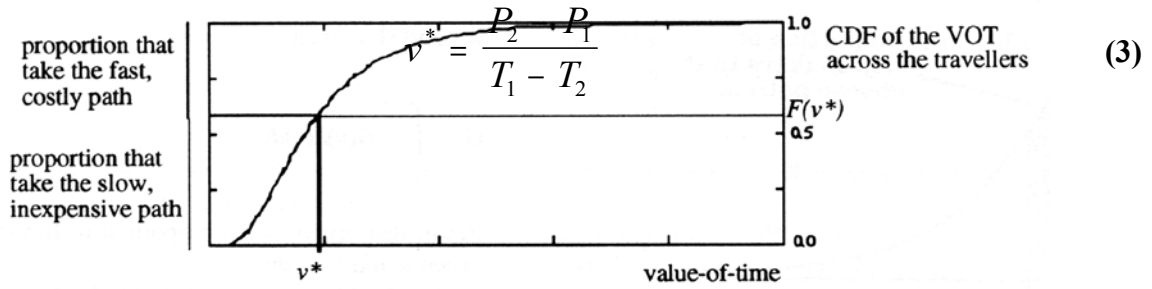


FIGURE 1 Market shares of two alternative paths.

Travelers with VOTs of $v_i \leq v^*$ choose the slow, inexpensive Path 1, whereas travelers with VOT of $v_i > v^*$ choose the fast, costly Path 2.

Given the statistical distribution of VOT across the trip-makers' population from its cumulative probability density function (CDF)

$$F(v) = \int_0^v h(x) dx \quad (4)$$

where $h(x)$ is the probability density function of VOT, and the proportion of people with VOT between x and $x + dx$ is $h(x)dx$, then (Figure 1)

- The market share of the first slow but inexpensive path is $S_1 = F(v^*) = \int_0^{v^*} h(x) dx$, and
- The market share of the second fast but costly path is $S_2 = 1 - S_1 = \int_{v^*}^{+\infty} h(x) dx$.

A way to infer the VOT statistical distribution is to derive it from the income distribution, which is in general well fitted by a log-normal probability density function (PDF) (6,7). Figure 2 depicts such a distribution. Another suggestion (12) is to consider a gamma distribution, which leads to a similar shape.

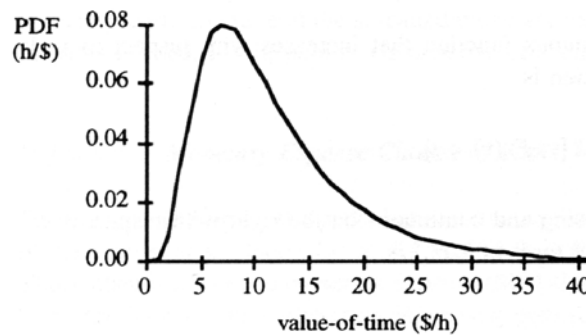


FIGURE 2 Log-normal distribution of VOT.

Efficient Paths

Call *efficient* a path such that there exists some positive VOT for which the path ensures a minimum generalized travel time. In the previous example there are only two paths, both of which are efficient. In most cases, however, there are numerous paths, among which only a few are efficient. If all paths are represented in a cost-versus-time diagram (where a path k is given coordinates T_k and P_k), the efficient paths are those with no alternative that would be both quicker and cheaper (Figure 3). In the cost-versus-time model only efficient paths may be assigned positive flows. If M efficient paths are ranked with respect to increasing prices, then the m th efficient path is traveled on by trip-makers with a VOT of v_i belonging to $[v_{m-1}^*, v_m^*]$, where v_m^* is the frontier VOT between efficient paths m and $m + 1$, defined as in Equation 3 as

$$v_m^* = \frac{P_{m+1} - P_m}{T_m - T_{m+1}} \quad (5)$$

Assuming a total trip rate of q , the m th efficient path is assigned a flow equal to

$$a \int_{v_{m-1}^*}^{v_m^*} h(v) dv = q[F(v_m^*) - F(v_{m-1}^*)] \quad (6)$$

Note that for consistency with respect to the first and last efficient alternatives, in the latter case the upper bound must be $+\infty$, and in the former case the lower bound must be 0. See Figure 4 for an illustration.

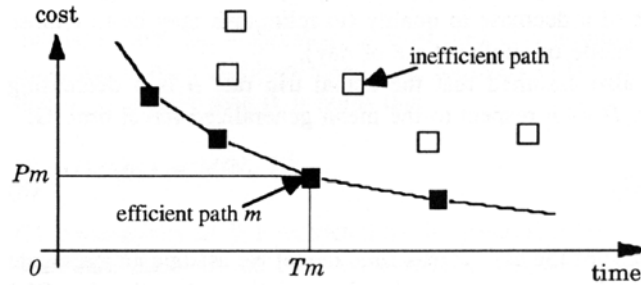


FIGURE 3 Efficient paths in a cost-versus-time diagram.

Contributions

The previous subsections have introduced the rule of sharing the traffic between the paths that underlies previous cost-versus-time models (6-8,12). To apply the rule those models have assumed that the prices and the travel times of the paths are fixed in advance.

However especially in urban road networks congestion effects may change the travel times of the paths and the definition of the set of efficient paths as well. Heuristic adaptations of the cost-versus-time sharing rule (10,11) have lacked a consistent theoretical framework.

The first contribution is also aimed at providing tools to take congestion into account within the cost-versus-time framework. The second contribution allows the volume of demand (the origin-destination traffic flow) to depend on the level of service.

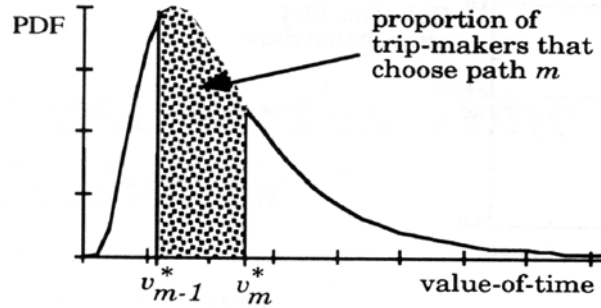


FIGURE 4 Market share of m th efficient path.

Congestion Effects

The more vehicles there are on a road the more delay each of them experiences. When modeling urban road networks it is necessary to allow for increasing travel time with respect to flow (13). Thus it is assumed that for each network arc a there is a travel time function $t_a = t_a(x_a)$ that relates travel time t_a to vehicular flow x_a . Defining the travel time of a path k as the sum of the travel times of the arcs a that make up path k , it thus depends on traffic flows.

Allowing for Elastic Demand

Elastic demand is the economic tool used to model the fact that a change in supply entails either more people making a trip if the change is an improvement or some people relinquishing a trip in the case of a decrease in quality (to relinquish may be to choose another mode or another time of day).

It is also assumed that the actual trip rate q is a decreasing function D with respect to the mean generalized travel time G :

$$q = D(G) \quad (8)$$

In the case of the cost-versus-time model we assume an aggregate measure of the mean generalized travel time; denoting by $G(v)$ the minimum generalized travel time experienced by a trip-maker with VOT v , then

$$G = \int_0^{+\infty} G(v)h(v)dv \quad (9)$$

Recall that $h(v)dv$ is the proportion of trip-makers with VOT between v and $v + dv$.

The elasticity of demand could also be modeled in a disaggregate way (14), but it would involve a mathematical framework with Hilbertian spaces of infinite dimension. Even more sophisticated is the Matisse model (9), which allows for cross-elasticities between segments of demand.

MATHEMATICAL DEVELOPMENTS

This section is rather technical. First, some notation is introduced. Second, *monetary expense classes of paths* that aggregate paths with the same price and that are the rigorous tool of dealing with the efficient paths are defined. Third, conditions that characterize a cost-versus-time equilibrium are set up. Fourth, a mathematical convex minimization program is presented; in that program the Kuhn-Tucker conditions are equivalent to the definition of a cost-versus-time equilibrium. Lastly existence and uniqueness are asserted without proof. For detailed proofs of the mathematical results, the reader is referred to previous reports (15,16).

Basic Notation

Demand Side

The demand is a set of couples $[D_{rs}(t), h_{rs}(v)]_{rs}$, where $r-s$ is an origin-destination (O-D) pair, $D_{rs}(t)$ is the demand function for trips between r and s (it is assumed to be a continuous and monotonically decreasing function with respect to the travel time t), v is a VOT, a number that belongs to a subset Ω of \mathbf{R}^+ , and $h_{rs}(v)$ is the probability density function of the random variable VOT among the travelers on O-D pair $r-s$; it is assumed to be a continuous and bounded function that remains nonnegative on the interior of its support. The cumulative density function associated with h_{rs} is as follows:

$$F_{rs}(v) = \int_0^v h_{rs}(\theta) d\theta$$

It is a continuous function that increases with respect to v . Its inverse function is

$$F_{rs}^{-1}(t) = \text{MAX}[v; F_{rs}(v) < t],$$

and is increasing and continuous on the right with respect to t .

A primitive form of $1/F_{rs}^{-1}$ is

$$E_{rs}(x) = \int_0^x 1 / F_{rs}^{-1}(\theta) d\theta$$

The inverse demand function

$$D_{rs}^{-1}(q) = \text{MAX}[t; D_{rs}(t) > q]$$

is decreasing and continuous on the right with respect to q .

q_{rs} is the trip rate from origin r to destination s .

Supply Side

a is an arc. x_a is the flow on arc a . $t_a(x_a)$ is the travel time function on arc a ; it is assumed to be a positive, continuous, and increasing function with respect to x_a .

Evaluation of Alternatives

k is a path from origin r to destination s , and it is assumed that it does not comprise any given arc more than once, f_k^{rs} is the flow on path k connecting O-D pair r - s , δ_{rs}^{ak} is the indicator variable ($\delta_{rs}^{ak} = 1$ if arc a is on route k between r and s and 0 otherwise), T_{rs}^k is the travel time proper on path k from r to s [$T_{rs}^k = \sum_a \delta_{rs}^{ak} t_a(x_a)$], P_{rs}^k is the monetary expense of path k from r to s (assumed to be nonnegative), and $G_{rs}^k(v)$ is the generalized travel time on path k from r to s as experienced by a trip-maker with VOT v of

$$G_{rs}^k(v) = T_{rs}^k + \frac{P_{rs}^k}{v} \quad (10)$$

Feasible Flow Pattern and Monetary Expense Classes of Paths

Definition 1: Feasible Flow Pattern

A feasible flow pattern is defined as a path flow vector $f=(f_{rs}^k)$ such that

$$\begin{aligned} - \forall r, s, k \quad f_{rs}^k &\geq 0 \\ - \forall r, s \quad q_{rs} &= \sum_k f_{rs}^k \leq q_{rso} \end{aligned}$$

where q_{rso} is some positive constant (any O-D flow is physically bounded).

The basic principle is to aggregate paths that share the same monetary cost. To that end the so-called monetary expense classes of paths is used. It will help to characterize efficient paths.

Definition 2: Monetary Expense Classes of Paths

For every O-D pair $r-s$ the paths between the equivalency classes of the equivalency relationship $R_{rs}[(kR_{rs}^l)]$ if $P_{rs}^k = P_{rs}^l$ are shared. Those classes are called *monetary expense* (ME) classes of routes. They are indexed with respect to increasing prices, with indexes from 1 to K_{rs} .

$I_{rs}(k)$ is defined as the class index of the path $(r-s)-k$ and Δ_{rs}^{ki} as an indicator variable: Δ_{rs}^{ki} is equal to 1 if i is equal to $I_{rs}(k)$ and Δ_{rs}^{ki} is equal to 0 otherwise.

Additional Notation Related to ME Classes of Paths

q_{rs}^m is equal to $\sum_k \Delta_{rs}^{km} f_{rs}^k$ is the traffic flow on the paths of the m th ME class from r to s , and Q_{rs}^m is equal to $\sum_{i \leq m} q_{rs}^i$ the traffic flow from r to s on the paths whose prices are less than or equal to the price on the paths of the m th class. It also holds that q_{rs} is equal to $\sum_m q_{rs}^m$ which is equal to $Q_{rs}^{K_{rs}} \bullet Q_{rs}^0$, is defined as 0 for ease of writing. P_{rs}^m denotes the monetary cost on the paths of the m th class for O-D pair $r-s$, and T_{rs}^m is the minimum travel time proper across the paths of this class.

The minimum generalized travel time experienced by a traveler with VOT v on the paths of the m th monetary expense class is

$$T_{rs}^m + P_{rs}^m / v \quad (11)$$

Cost-Versus-Time Equilibrium Conditions

Definition 3: Cost-Versus-Time Equilibrium Conditions

The feasible flow pattern f is a cost-versus-time equilibrium if and only if the following conditions (C1 to C3) are satisfied:

$$C1: \forall r, s, k \quad f_{rs}^k > 0 \Rightarrow T_{rs}^k = T_{rs}^{Irs(k)} \quad (12)$$

For every O-D pair $r-s$, for two monetary expenses classes m and n that are utilized ($q_{rs}^m > 0$ and $q_{rs}^n > 0$), it holds that

$$C2: T_{rs}^m + \sum_{j=m}^{K_{rs}-1} \frac{P_{rs}^j - P_{rs}^{j+1}}{F_{rs}^{-1}(Q_{rs}^j / q_{rs})} = T_{rs}^n + \sum_{j=n}^{K_{rs}-1} \frac{P_{rs}^j - P_{rs}^{j+1}}{F_{rs}^{-1}(Q_{rs}^j / q_{rs})} \quad (13)$$

In the variable-demand case for every O-D pair $r-s$ such that q_{rs} greater than 0, it holds that

$$C3: D_{rs}^{-1}(q_{rs}) = \sum_{m=1}^{K_{rs}} \left[\frac{q_{rs}^m}{q_{rs}} T_{rs}^m + P_{rs}^m \left(E_{rs} \left(\frac{Q_{rs}^m}{q_{rs}} \right) - E_{rs} \left(\frac{Q_{rs}^{m-1}}{q_{rs}} \right) \right) \right] \quad (14)$$

Economic Interpretation

The equilibrium conditions may be compared with the definitional conditions of a Wardropian user equilibrium (W1 and W2):

$$W1: \forall r, s, k \quad f_{rs}^k > 0 \Rightarrow T_{rs}^k = \text{MIN}_k T_{rs}^k$$

that is, a path that is traveled on must present a minimum travel time, and in the variable-demand case, for every O-D pair r - s such that q_{rs} is greater than 0, it holds that

$$W2: D_{rs}^{-1}(q_{rs}) = \text{MIN}_k T_{rs}^k$$

C1 corresponds to W1 restricted to the paths that belong to the same ME class; in the cost-versus-time model, the equilibration of flows owing to congestion effects prevails only *inside* each of the ME classes of paths.

C3 is analogous to W2 since it relates the volume of demand to a mean minimum generalized travel time. To see that the term on the right side of C3 stands for the definition of the generalized travel time presented in Equation 8, one must change the variables under the integration symbol:

$$\begin{aligned} \sum_{i=1}^{K_{rs}} T_{rs}^i \frac{q_{rs}^i}{q_{rs}} + P_{rs}^i \left[E_{rs} \left(\frac{Q_{rs}^j}{q_{rs}} \right) - E_{rs} \left(\frac{Q_{rs}^{i-1}}{q_{rs}} \right) \right] = \\ \sum_i \int_{F_{rs}^{-1}(Q_{rs}^{i-1}/q_{rs})}^{F_{rs}^{-1}(Q_{rs}^i/q_{rs})} (T_{rs}^i + P_{rs}^i / v) h_{rs}(v) dv \geq \sum_i \int_{F_{rs}^{-1}(Q_{rs}^{i-1}/q_{rs})}^{F_{rs}^{-1}(Q_{rs}^i/q_{rs})} \text{MIN}_j (T_{rs}^i + P_{rs}^i / v) h_{rs}(v) dv = \\ \int_{\Omega} \text{MIN}_j (T_{rs}^i + P_{rs}^i / v) h_{rs}(v) dv \end{aligned}$$

with equality being the case when C1 and C2 are satisfied.

C2 is specific to the cost-versus-time model. It determines the market share of each ME class of paths. If the ME classes i and $(n = i + p)$ are utilized when the classes $i + j$ for j in $[1; p - 1]$ are not (that is, $Q_{rs}^{i+j} = Q_{rs}^i$), then C2 reduces to

$$F_{rs}^{-1}(Q_{rs}^i / q_{rs}) = (P_{rs}^{j+p} - P_{rs}^i) / T_{rs}^i - T_{rs}^{i+p} \quad (15)$$

$F_{rs}^{-1}(Q_{rs}^i / q_{rs})$ is the frontier VOT between the alternatives i and $i + p$; when C2 holds the paths of the i th class are used by the travelers whose VOT belongs to $[v_{rs,i-1}^*, v_{rs,i}^*]$ because these paths enable them to minimize $\mathbf{T} + \mathbf{P}/v$ (compare Equation 15 with Equation 5).

Extreme Characterization of a Cost-Versus-Time Equilibrium

Theorem 1 is the convex program for the cost-versus-time equilibrium. The feasible flow pattern \mathbf{f} is a cost-versus-time equilibrium if and only if it solves the extremal convex problem $MIN J(\mathbf{f})$ on the set of all feasible flow patterns, where function J is defined as

$$J(\mathbf{f}) = \sum_a \int_0^{x_a} t_a(x) dx + \sum_{rs} \left(\left\{ q_{rs} \sum_{j=1}^{K_{rs}} P_{rs}^j \left[E_{rs} \left(\frac{Q_{rs}^j}{q_{rs}} \right) - E_{rs} \left(\frac{Q_{rs}^{j-1}}{q_{rs}} \right) \right] \right\} - \int_0^{q_{rs}} D_{rs}^{-1}(q) dq \right) \quad (16)$$

subject to the definitional constraints

$$x_a = \sum_{rsk} \delta_{rs}^{ak} f_{rs}^k \quad (16a)$$

$$q_{rs}^i = \sum_k \Delta_{rs}^{ki} f_{rs}^k \quad (16b)$$

$$Q_{rs}^i = \sum_{j \leq i} q_{rs}^j \quad (16c)$$

$$q_{rs} = \sum_{j=1}^{K_{rs}} q_{rs}^j \quad (16d)$$

and the nonnegativity constraints

$$f_{rs}^k \geq 0 \quad (16e)$$

The first sum in the definition of J refers to the travel times proper that people try to minimize. The second one is related to the MEs that people also try to minimize. The third one is close to the opposite of a consumers' surplus.

Existence and Uniqueness of Equilibrium

Theorem 2: Existence

There exists at least one cost-versus-time equilibrium.

Theorem 3: About Uniqueness

If the travel time functions $t_a(x_a)$ are strictly increasing then at equilibrium the arc flows are unique, as are the frontier VOTS. In the fixed-demand case the flow on each ME class of the paths is unique. In the variable-demand case if the demand functions D_r are strictly decreasing, then the trip rates as well as the flows on each ME expense class of paths are unique.

MSA ALGORITHM

Because practitioners are mainly interested in convenient, robust methods, an algorithm that avoids path storage and enumeration is presented. It is a Monte-Carlo method based on random simulation, as will be discussed further.

Two assumptions help to simplify the procedure:

- The price of path k depends only on the prices m_a of the arcs a that are traveled on:

$$P_{rs}^k = \sum_a \delta_{rs}^{a,k} m_a$$

- All O-D pairs with the same origin have the same PDF for VOT to avoid computing the shortest paths for each O-D pair.

Procedure

Step 0: Initialization

- Set iteration counter $n = 0$.
- Choose a sequence α_k of real numbers such that $(0 \leq \alpha_k \leq 1)$, $(\sum \alpha_k = \infty)$ and $(\sum \alpha_k^2 < \infty)$.
- Find an initial feasible flow pattern $[x_a^{(0)}; q_{rs}^{(0)}]$ for the variable-demand case only].

In the fixed-demand case the initial feasible flow pattern may be obtained through an all-or-nothing assignment on the basis of times $[t_a(0)]$.

In the variable-demand case the free-flow pattern may be used as an initial flow pattern; set O-D generalized travel time variables $G_{rs}^{(0)}$ to the most realistic available value (from past assignments or some point of the demand curve that corresponds to a realistic mean generalized time).

Step 1: Arc Travel Time Update

- Set $n = n + 1$.
- Set $t_a^{(n)} = [t_a(x_a^{(n)})]$.

Step 2: Direction Finding

- For each origin r select by random sampling a VOT $v_r^{(n)}$. Compute the shortest paths to all destinations s on the basis of the arc generalized travel times $[t_a^{(n)} + m_a/v_r^{(n)}]$, yielding auxiliary O-D generalized travel times $G1_{rs}^{(n)}$. For each destination s assign an auxiliary O-D flow $q1_{rs}^{(n)}$ on the shortest path thus determined, in which $q1_{rs}^{(n)}$ is equal either to q_{rs} in the fixed-demand case or to $D_{rs}[G_{rs}^{(n)}]$ in the variable-demand case.

Assignment of traffic of all O-D pairs yields an auxiliary arc flow pattern $x1_a^{(n)}$.

Step 3: Arc Flow (and O-D Time) Update

- Set $x_a^{(n+1)} = x_a^{(n)} + \alpha_n (x1_a^{(n)} - x_a^{(n)})$.
- In the variable-demand case set $G_{rs}^{(n+1)} = G_{rs}^{(n)} + \alpha_n [G1_{rs}^{(n)} - G_{rs}^{(n)}]$.

Step 4: Convergence Criterion

- Apply a convergence test, either a maximum number of iterations or a test on the maximum value (among the arcs a of the network) of the change in $\sum_{k=1}^n \alpha_k \cdot x_a^{(k)} / \sum_{k=1}^n \alpha_k$ from the previous iteration $n - 1$ to the current one, n . If the test is satisfied, then terminate or go to Step 1.

Comments

The suggested algorithm is a twice-streamlined implementation of the method of successive averages (17,18) (with regard to streamlined algorithms).

- Step 2 begins with a random sampling of the VOT-, if it was iterated many times the cumulative mean of the auxiliary flow pattern thus obtained would yield a descent direction for J at the current point of Step 1. The streamlined algorithm with one single internal sampling in Step 2 provides the best efficiency (18).

- In the variable-demand case the second part of Step 3 is a further streamlining that allows one to compute the minimum generalized travel time without storing paths.

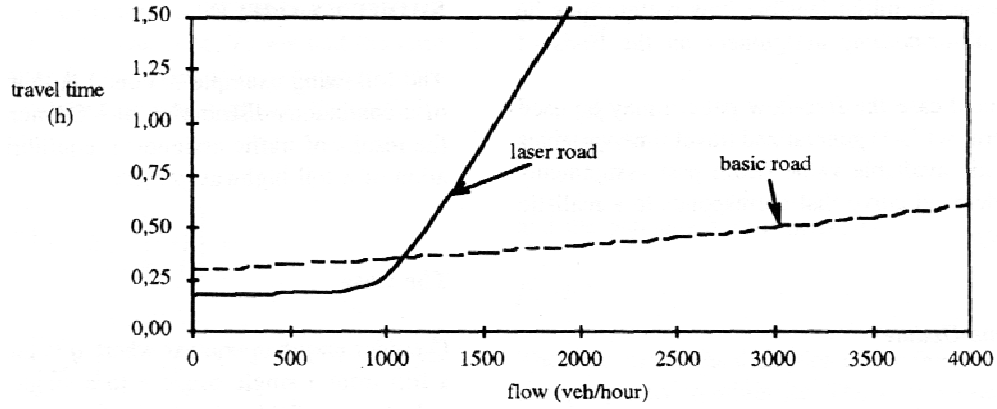


FIGURE 5 Travel time functions for two routes.

SHORT EXAMPLE

The following example is aimed at demonstrating that modeling of a continuous-distribution of VOT across travelers may change the results of traffic assignment equilibrium models in the evaluation of a toll highway project.

The Case

Consider an urban context where q vehicles per hour are to make a trip from a single origin r to a single destination s . There are only two available routes, the first one (the basic road) being a free route on the regular urban network and the second one being a toll-charged route designed to allow for quick traveling without congestion. The latter route is named the laser road, from the original idea of GTM (19) to build underground, passenger car-only toll motorways in Paris.

Supply Side

Assume that the generalized time on road a for a traveler with VOT is

$$G_a(x_a, v) = t_a(x_a) + \frac{P_a}{v} =$$

$$T_a \cdot [1 + \gamma_a + \sqrt{\alpha_a^2 \cdot (1 - x_a / N_a)^2 + \beta_a^2} - \alpha_a \cdot (1 - x_a / N_a) - \beta_a] + \frac{P_a}{v}$$

where

P_a = toll fare on road a ;
 N_a = measure of *practical* capacity (i.e., the traffic flow at which point the service level on the arc decreases sharply); and
 $\alpha_a, \beta_a, \gamma_a$ = parameters to model the effects of congestion such that (Figure 5):

The values of the parameters are as follows. For the laser arc T_a equals 0.18 hr, N_a equals 1,000 vehicles/hr, α_a equals 4.0, and γ_a equals 0.5. For the basic arc T_a equals 0.30 hr, N_a equals 5,000 vehicles/hr, α_a equals 2.5, and γ_a equals 1.5.

Demand Side

Assume that the VOT is distributed according to a log-normal PDF as depicted in Figure 2. The log-normal PDF is characterized by its median value of \$10/hr, and the standard deviation of its natural logarithm is set equal to 0.6 (20). The total trip rate q is fixed to 3,000 vehicles/hr.

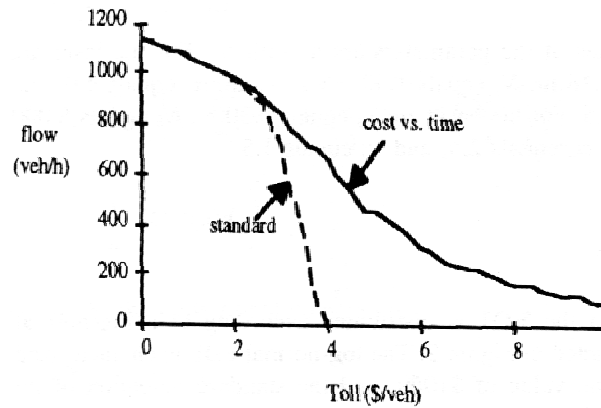


FIGURE 6 Traffic on toll road (laser route).

Numerical Evidence and Discussion

Calculate the traffic on the toll road and the toll revenues as functions of the toll asked for on the laser road by two different models: a cost-versus-time model and a standard model in which all travelers have the same aggregate VOT (the mean of the VOT distribution in the cost-versus-time model).

As soon as the toll is high enough to significantly differentiate the two routes the drawbacks of the standard, single-VOT model appear; it is unable to calculate either the optimal level of fare or the maximum revenues that are yielded by the more realistic cost-versus-time model.

Furthermore the standard model does not yield robust results; the fare that gives the maximum revenue is very close to another fare at which nobody travels on the toll route (Figures 6 and 7).

CONCLUSION

Cost-versus-time equilibrium assignment, as any assignment over a network, deals with a demand and a supply that are at odds with each other. It has been defined as a double equilibrium:

- An equilibrium between supply and demand, and
- A Pareto equilibrium between suppliers (the paths).

The mathematic and algorithmic tools that enable computation of this equilibrium are especially useful for the evaluation of toll highway projects in urban contexts.

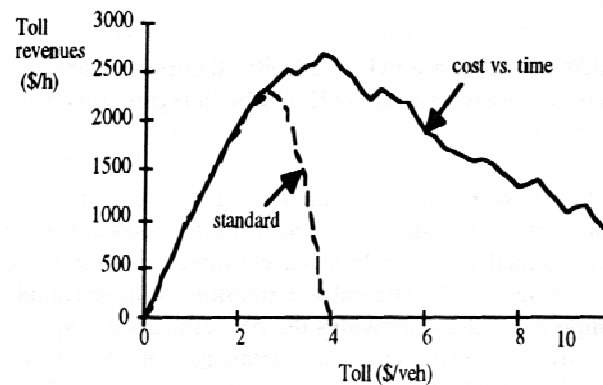


FIGURE 7 Toll revenues.

The cost-versus-time model does not invalidate the advantages of the multiple user classes model, which faithfully represents the interaction on the supply side, notably different types of vehicles (e.g., with respect to size or passenger car unit equivalents).

The model described here is primarily demand related. The continuous distribution of the VOT gives robustness to the assignment and also enables one to test sensitivity to parameters like the mean value or the standard deviation of the VOT distribution function. An obvious extension is to introduce several classes of vehicles, each one with a continuously distributed VOT.

A truly disaggregate model requires that nonuniform multicriteria measures of the generalized cost (or of the satisfaction) in the demand be taken into account. The cost-versus-time model is a significant step in that direction for traffic assignment. Although close to the stochastic equilibrium models with respect to the algorithm introduced here, the economic background is quite different, focusing on explaining the deterministic part of the utility function rather than calibrating its random component like current stochastic assignment models do.

From a mathematical point of view researchers are provided with a computationally tractable model that extends the model of Beckmann et al. (13). Apart from the method presented here, deterministic algorithms are also available (21).

ACKNOWLEDGMENTS

The author is indebted to colleagues at the Institut National de Recherche sur les Transports et leur Sécurité, Roger Marche and Francois Barbier Saint-Hilaire, for their seminal work about cost-versus-

time models and Olivier Morellet for many discussions about the Wardropian equilibrium. The author benefited from the comments of Michael Florian and Ennio Cascetta at an early stage of the work. Helpful editorial suggestions from anonymous referees and Laura Marmorstein are gratefully acknowledged.

REFERENCES

1. Dial, R. B. A Probabilistic Multipath Traffic Assignment Model Which Obviates Path Enumeration. *Transportation Research*, Vol. 5, 1971, pp. 83-111.
2. Daganzo, C. F., and Y Sheffi. On Stochastic Models of Traffic Assignment. *Transportation Science*, Vol. 11, No. 3, 1977, pp. 253-274.
3. Dafermos, S. C. The Traffic Assignment Problem for Multiclass-User Transportation Networks. *Transportation Science*, Vol. 6, 1972, pp. 73-87.
4. Daganzo, C. F. Stochastic Network Equilibrium with Multiple Vehicle Types and Asymmetric, Indefinite Link Cost Jacobians. *Transportation Science*, Vol. 17, 1983, pp. 282-300.
5. Thomas, R. *Traffic Assignment Techniques*. Avebury Technical, Aldershot, England, 1991.
6. Marche, R. *Le Mod le TRIP*. SETEC, Paris, 1973.
7. Marche, R. Pour mieux Comprendre les Déplacements Interrégionaux de Voyageurs: UN Modèle Multimodal de Demande. Description du Modèle et des Résultats. *Les Cahiers Scientifiques de la Revue Transports* 3, 1980, pp. 52-68.
8. Abraham, C., and J. D. Blanchet. Le Modèle Prix-Temps. *Revue de l'Aviation Civile*, 1973.
9. Julien, H., and O. Morellet. *MATISSE, un Mod le Int grant étroitement Induction et Portage Modal Fin du Trafic*. INRETS Report 129. INRETS, Arcueil, France, 1990.
10. Barbier Saint-Hilaire, F. *Syst me DAVIS Equilibre*. Internal Reports. INRETS, Arcueil, France, 1993.
11. Daiko, T., T. Okubo, H. Moritsu, and T. Morikawa. *Practical Traffic Assignment for Multiple Highway Routes Using Distribution of Values of Time*. Paper presented at the Sixth WCTR, Lyon, France, 1992.
12. Dial, R. B. A Model and Algorithm for Multicriteria Route-Mode Choice. *Transportation Research B*, Vol. 13, 1979, pp. 311-316.
13. Beckmann, M., C. B. McGuire, and C. B. Winsten. *Studies in the Economics of Transportation*. Yale University Press, New Haven, Conn., 1956.
14. Leurent, F. Modelling Elastic, Disaggregate Demand. *Proc. First Meeting of the Euro Working Group on Urban Traffic and Transportation*. Technical University of Munich, Munich, Germany, 1993.
15. Leurent, R. NT 91.3 Affectation Prix-Temps sur un Réseau: Formulation Extrême et Algorithme pour un Cas Simple. Working paper. INRETS, Arcueil, France, 1991.

16. Leurent, F. Cost Versus Time Equilibrium over a Network. *EJOR*, Vol. 71, No. 2, 1993, pp. 205-221.
17. Powell, W B., and Y Sheffi. The Convergence of Equilibrium Algorithms with Predetermined Step Sizes. *Transportation Science*, Vol. 16, No. 1, 1982, pp. 45-55.
18. Sheffi, Y *Urban Transportation Networks*. Prentice-Hall, Incorporated, Englewoods Cliffs, N.J., 1984.
19. GTM. *Projet LASER, Documents Techniques de Présentation*. Groupe GTM-Entrepose, Nanterre, France, 1988.
20. Marche, R. *Notes LASER 88.1 à 88.7*. Working papers. INRETS, Arcueil, France, 1988.
21. Leurent, F. MEDOC, *Modèle d'équilibre entre Demande et et Offre de Circulation*. Working paper. INRETS, Arcueil, France, 1993.

Publication of this paper sponsored by Committee on Transportation Supply Analysis.

TRAFFIC ASSIGNMENT UNDER ENVIRONMENTAL AND EQUITY OBJECTIVES

LAURENCE R. RILETT AND CHRISTINE M. BENEDEK¹

Two recent changes in the transportation field may have a profound effect on traffic assignment techniques. The first is the increasing importance of environmental objectives, such as reducing air pollution, within the policies of traffic system authorities. The second change is the advent of the intelligent vehicle-highway system (IVHS), which, among other attributes, has the potential to be used to implement new methods of controlling vehicular emissions. The fact that historic traffic assignment techniques may be inadequate for modeling the traffic systems that will operate under IVHS with environmental objectives--primarily when traffic follows routes that are based on equitable rather than equilibrium or optimal considerations--is illustrated. Then it is shown that when IVHS policies that attempt to reduce system travel time are implemented, other objectives such as reducing environmental pollution may actually increase. A network from Ottawa, Ontario, Canada, is used as a test bed.

A number of objectives are generally associated with the proposed intelligent vehicle-highway system (IVHS). One of the most commonly cited goals is to reduce urban traffic congestion with a corresponding reduction in average trip travel time. This has been an ongoing objective of transportation authorities since the first urban traffic networks were constructed. Previously the most common method of achieving this goal was to increase capacity through building infrastructure. Another objective that is often cited is to reduce negative transportation by-products such as noise and air pollution. This has become an increasingly important goal over the last 30 years mainly because of increased public awareness of the dangers of pollution and a public willingness to reduce this pollution. Historically the primary means of reducing air pollution have been through legislated emission standards on vehicles.

This paper examines the implications on traffic networks of using the recently proposed IVHS such as the advanced traffic management system (ATMS) and the advanced traveler information system (ATIS) to achieve the goals stated above. These proposed systems may be used to achieve the objectives in either an active or a passive manner. The former would entail such things as a centralized route guidance system (RGS) in which vehicles are explicitly, given the routes that they must follow, whereas an example of the latter would be an electronic toll collection system in which drivers are free to choose their own routes but are charged for their use of the road or the amount of pollution that then, produce. Note that in both the passive and the active systems different goals or combination of goals may be used.

The second section illustrates the need for new traffic assignment techniques that better represent the shift toward environmental objectives that has recently taken place. In the third section assignment techniques based on environmental and equitable objectives are examined on a two-node network to illustrate the concepts. This is followed by a sensitivity analysis of traffic assignment based on environmental objectives on a network from Ottawa, Ontario, Canada, to

¹ Department of Civil Engineering, University of Alberta, Edmonton, Alberta T6G 2G7, Canada.

identify any trends on realistic networks and any potential problems in using traditional assignment procedures.

RECENT DEVELOPMENTS IN TRAFFIC ENGINEERING

Two major developments in recent years are forcing traffic engineers to reexamine the techniques and objectives of traffic assignment. The first shift is the rapid advancement in IVHS technologies, particularly in-vehicle RGSs, in which it is at least theoretically possible that drivers may be explicitly routed through the network on the basis of the routes that are calculated external to the driver or the vehicle. At a minimum IVHS technologies will influence driver route selection by providing timely information on the state of the network. For example the use of automatic toll collection on the road network or the use of changeable message signs could change the route selection process of drivers by changing the perceived attributes of competing routes.

It is often assumed that because traffic assignment is based on the concept of generalized cost the traditional assignment techniques will be applicable for analyzing IVHS. The major changes required in the traditional procedures include modeling multiple user classes (RGS and non-RGS) and modeling dynamic traffic assignment (1,2). Although very complex, these topics will not be examined here because the main purpose of this paper is to illustrate potential problems in traffic networks when different objectives are used and to illustrate the need for assignment techniques that can model equitable as opposed to equilibrium or optimal assignments. It will be assumed in all of the analyses in this paper that all drivers have the same attributes and the same access to information. Therefore user equilibrium (UE) techniques will be used for modeling IVHS in which the drivers select their routes on the basis of their own objectives, and system optimal (SO) techniques will be used for modeling IVHS in which the routes are explicitly sent to drivers and are based on system considerations.

The second change that will affect traffic assignment techniques is related to the prominent role that environmental issues have recently played in transportation project decisions, in particular the significant interest that has recently been expressed concerning the consequences of vehicular emissions. Reducing vehicular emissions has been an ongoing goal of many authorities over the past 20 years, with a number of U.S. states and Canadian provinces instituting relatively stringent pollution control programs. These programs may be defined as passive in nature, in that most regulate the emission levels from the vehicles. However the total amount of pollutant emitted by a vehicle is not regulated, and consequently there is little incentive for individual users to reduce pollution. With the advent of IVHS it is now recognized by many traffic authorities that more active measures may be used to reduce pollution. As an example it may be decided to use a centralized RGS to directly route vehicles so as to minimize air pollutant emissions during particular periods of the day or in particular locations. A more passive and realistic example would involve charging drivers on the basis of the amount of pollution that they produce (and where that pollution is produced) in the hope of reducing emissions.

Because traffic assignment techniques are based on the concept of generalized cost it may be assumed that traditional assignment techniques will also be applicable for analyzing route selection on the basis of environmental impacts. The only changes required to implement the assignments discussed above would be the development of appropriate generalized cost

functions. However identification of the generalized cost function that would be used is problematic. There has been little research into which of the relevant factors (i.e., noise pollution, fuel consumption, etc.) should be included in the generalized cost function or what the relative weights for each of the relative factors should be. Regardless of which components are used in the generalized cost function it is important to note that in the context of environmental concerns the process and objectives of traffic assignment shift from factors that solely concern the individual drivers or the system operators to factors that also consider the effects on individual segments of society. That is it may not be enough to say that the individual drivers or the transportation network operators will benefit as a result of the implementation of IVHS technologies but rather that no segment of society will be unduly affected in a negative manner. For example a political decision may be made that the reduction of negative transportation by-products should be a major policy objective regardless of the impact on individual drivers. The traffic assignment techniques will have to reflect this new reality if meaningful analyses of environmental objectives and IVHS implementation are to be studied. Consequently it is not clear that the objectives of the SO or UE traffic assignment, even with an appropriate generalized cost function, will be adequate for analyzing the assignment of traffic under these new conditions.

It is useful at this point to examine (a) how the concept of equity and environmental concerns may influence the actual route selection process of the drivers and (b) how these changes may be modeled by using traffic assignment procedures.

Consider a negative product X , where X may be the noise pollution, air pollutants, and so on that are caused by vehicular traffic. It may be decided that the amount of X produced should be controlled through the use of an IVHS. The following is a discussion of the different objectives that may be chosen and the strategies that may be employed to meet the objective of reducing X . Also included are potential means of modeling these strategies in traffic assignment procedures.

It may be decided that the objective of the IVHS is to minimize the total amount of X produced. The objective could be achieved by giving explicit routes to the individual vehicles through a centralized RGS. The traffic assignment could be modeled by using the SO concepts discussed previously in which the generalized cost is a function only of X rather than of travel time.

Alternatively it may be decided that although decrease in the production of X is the primary objective, it would be better (i.e., politically better) to charge users on the basis of their production of X and let the drivers decide their own routes. In the real network an electronic toll system in which the drivers are charged on the amount of X that they are responsible for producing would be set up. This is directly analogous to "occasional cost" pricing, whereby consumers (drivers) pay only for what they directly consume. This type of system could be modeled by using a standard UE traffic assignment, with X being the sole parameter in the generalized cost function.

Both of the scenarios presented above assume that the assignment of vehicles will be based on the needs of the individual drivers or those of the system as a whole. However it is not unreasonable to assume that society will also wish to minimize the amount of X produced on particular segments of the population. The following two sections will illustrate two equity concepts that could be used in traffic assignment.

As an example the people living near major roadways may wish the vehicles to be routed through the network such that the total amount of X released on their streets does not exceed some maximum safety standard (i.e., for health reasons). This would correspond to traditional assignment techniques that have an explicit link capacity constraint (3) in which the link capacity is not a function of the amount of vehicles on the link but rather the cumulative amount of pollutant X that the vehicles produce on the link. Depending on the method chosen by the authorities for achieving the objective a UE or SO traffic assignment heuristic procedure could be used to model the process. However note that secondary objectives, such as minimizing the number of homes exposed to relatively high levels of X , may also be employed, and consequently new assignment techniques could be required.

Last the vehicles may be assigned to a street network in such a way as to ensure that the amount of X released on all streets (or a subset of streets) is the same. Under this system equitable (SE) scenario vehicles are routed through the network (on the basis of the routes directly broadcast to the vehicles) such that no one group of people living near the traffic network is affected more than any other group of people. This may at first seem to be an extreme example, but there are currently a number of situations in which traffic control devices are operated such that the negative externalities of traffic (i.e., noise) are "distributed" as evenly as possible among competing routes.

It should be pointed out that although there is a wide range of equity definitions (4) in this paper, only the SE concept defined above will be used. It is also important to stress the fact that in the preceding two types of assignments it is the objectives of the people living near the roadway and not those of the individual drivers or system operators that are of paramount importance.

TRAFFIC ASSIGNMENT BASED ON ENVIRONMENTAL AND EQUITY OBJECTIVES

Until recently the above differentiation and the following examples would be of more academic rather than practical interest. The advent of the in-vehicle RGS, however, has created the potential to change people's route selection behaviors, either directly (explicit directions) or indirectly (variable user charges). Given the demand by the public to reduce pollution it is very reasonable to assume that environmental objectives may be increasingly important in the traffic assignment process. The following sections will examine the potential effects that different objectives could have with a sample test network and a network representing Ottawa, Ontario, Canada, serving as examples.

Carbon Monoxide (CO) Emissions

There are a number of fuel consumption-pollutant emission models of various complexities. For the analyses performed in this paper a macroscopic relationship used in the TRANSYT 7-F model was adopted (5). The general function of the model is

$$ROP = \frac{Ae^{Bv}}{Cv} \quad (1)$$

where

ROP = rate of production [fuel (gal-vehicle/ft) or pollutant (g-vehicle/ft)],

v = average vehicular velocity on link (ft/sec), and

A, B, and C = constants.

It is assumed that the velocities of the vehicles are constant along each link and the grades on all roads are 0 percent. The velocity on the link is derived by dividing the distance of the link by the travel time. The total amount of pollutant produced per vehicle on any given link is then calculated by multiplying the production rate by the distance of the link.

Equation 1 is applicable for estimating fuel consumption. CO emissions, hydrocarbon emissions, and nitrogen oxide emissions. It was decided that traffic assignment would be examined only on the basis of CO emission rates. There are two reasons for this. The first is that because of the similarity in the form of the production functions the assignment results obtained on the basis of all the pollutants would be similar. The second is that CO is generally considered one of the most critical pollutants where levels need to be reduced (6). The form of the CO production function used in the analysis is given in Equation 2:

$$ROP = 3.3963 \frac{e^{0.014561v}}{1,000v} \quad (2)$$

where ROP is the rate of production of CO (g-vehicle/ft), and v is the average vehicular velocity on link (ft/sec).

Traffic Assignment with Environmental Objectives: Sample Network

To examine the concepts discussed above it is first useful to examine the changes in link flows when assigning the vehicles on the basis of travel time and CO emissions for an example network. The sample network consists of two links and two nodes, as illustrated in Figure 1. The origin-destination (O-D) demand consists of 8,000 vehicles/hr that travel from node 1 to node 2. There are two potential routes for these vehicles. The first, Route 1, is a two-lane freeway route that is 2000 m long, has a free-flow speed of 100 km/hr, and a capacity of 2,000 vehicles/hr/lane. The second, Route 2, is a shorter two-lane arterial route that is 1,000 m long, but with a lower free-flow travel speed of 60 km/hr and the same lane capacity as that of Route 1.

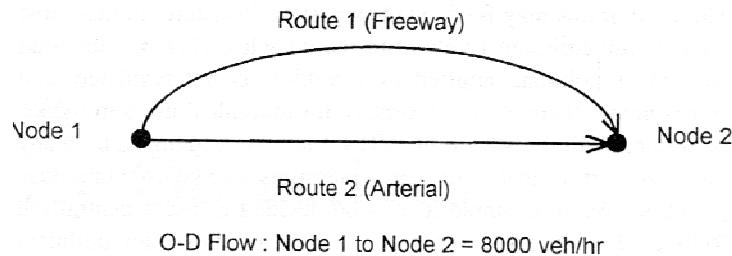


FIGURE 1 Sample network.

In the paper the acronym UE-TT refers to a user equilibrium traffic assignment based on travel time, whereas the acronym SO-TT refers to a system optimal traffic assignment based on travel time. Similarly UE-CO refers to a user equilibrium traffic assignment based on CO emissions, and SO-CO refers to a system optimal assignment based on CO emissions.

For the sample network a UE-TT assignment results in a flow on Route 1 of 5,090 vehicles/hr and a flow on Route 2 of 2,910 vehicles/hr. The travel time on both routes is 100.3 sec, and the total travel time on the network is 229.9 vehicle hr. If a traffic operations engineer was able to assign the vehicles to the networks to minimize total system travel time (SO-TT). The flow on Route 1 would increase to 5,218.2 vehicles/hr and the flow on Route 2 would decrease to 2,781.8 vehicles/hr. This would reduce the total system travel time to 222.1 hr.

Figure 2 illustrates the relationship between CO emissions on both routes of the sample problem as a function of flow on Route 1. It can be seen from Figure 2 that if the objective is to minimize the total CO emissions (SO-CO assignment) then 5,161 vehicles would take Route 1, which would result in 6.09 kg of CO being emitted into the atmosphere per hr. This is shown as point a on Figure 2. The rate of CO emitted by each of the vehicles on Route 1 is 0.88 g/vehicle, and on Route 2 it is 0.54 g/vehicle, which is a difference of approximately 40 percent. In total Route 1 receives 1.54 kg of CO, whereas Route 2 receives 4.55 kg, or approximately three times as much.

The difference between the flows from the SO-TT and UE-TT solutions and the SO-CO solution is on the order of 1 percent.

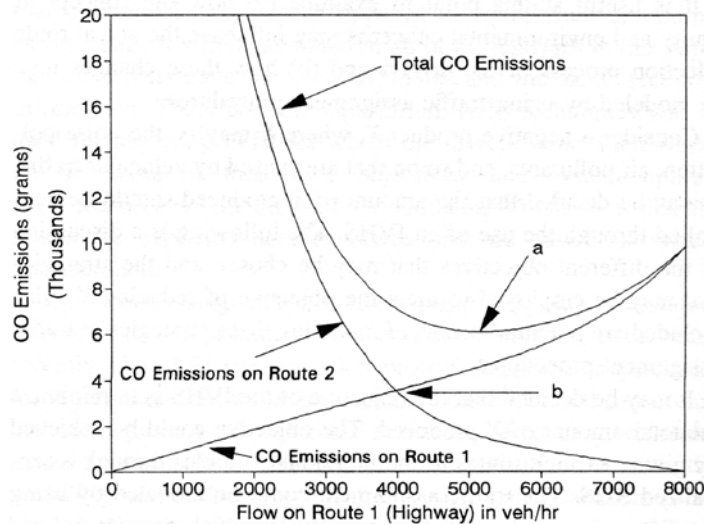


FIGURE 2 CO emissions versus volume on Route 1.

This is also confirmed by the fact that the travel time on Route 1 is 101.9 sec and on Route 2 it is 96.5 sec. Therefore for this simple example the UE-TT solution is roughly equivalent to an SO-CO assignment.

If the objective of the assignment is to ensure that both routes have equal CO emission levels (SE assignment) then 3,982 vehicles would be assigned to Route 1 and 4,018 vehicles would be

assigned to Route 2. This assignment is shown as point b on Figure 2. The difference in route flows between the SE and the SO-CO solutions is on the order of 25 percent. The change in route flows would increase the total CO emitted by 14 percent, to 7.10 kg/hr, in which each route experiences an emission rate of 3.55 kg/hr. The travel time on Route 1 is 82.6 sec, and the travel time on Route 2 is 206.5 sec. Therefore unless the drivers are explicitly assigned to the network in the proportions given above this would be an unstable solution.

Figure 3 illustrates the amount of CO produced per vehicle as a function of the flow on Route 1. If the vehicles were allowed to choose their own route but were charged for CO emissions (and considered only the cost of this in their route selection process) then 3,966 vehicles would take Route 1 and 4,034 vehicles would take Route 2. The UE-CO solution is illustrated by point a in Figure 3, in which it may be seen that the vehicles on both routes emit 0.89 g/vehicle. The travel time on Route 1 is 82.43 sec, and the travel time on Route 2 is 209 seconds.

It may be seen from the above analysis that the SE and UE solutions on the basis of CO emissions have similar results in terms of route volumes. The primary difference (aside from the objectives) between the two is that in the SE assignment it is assumed that the drivers have routes chosen for them whereas in the UE assignment the drivers select their own routes on the basis of the amount of CO they produce. This indicates that charging vehicles for pollutant emissions could achieve the same equitable environmental goals as routing them by using a centralized RGS. The negative side to this strategy is that charging for use of the road on the basis of environmental concerns could actually increase the total amount of CO produced.

It may also be seen that unlike the SO-TT and UE-TT solutions, the SO-CO and UE-CO solutions are significantly different. Therefore when IVHS strategies are implemented the objectives adopted could have a significant impact on the link flows and the amount of pollution produced. The following sections will examine whether these findings hold true for more realistic networks.

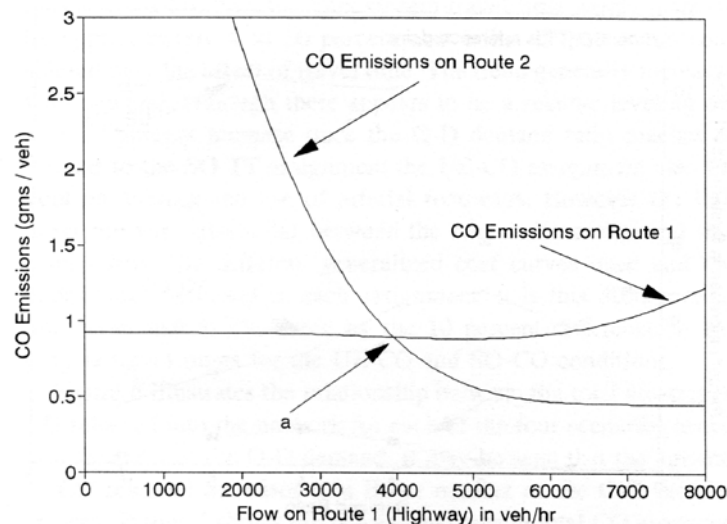


FIGURE 3 Route CO emission rates versus volume on Route 1.

Traffic Assignment with Environmental Objectives: Ottawa Network

A network from Ottawa, Ontario, Canada was chosen as a test bed to examine if the trends that were found for the simple example also exist in larger networks. The Ottawa network consists of 1,402 links, 646 nodes, and 67 zones. It is linear in shape, with the Queensway, the major highway in Ottawa, running in an eastwest direction through the center of the city. The assumptions and relationships that were used in the sample analysis are also used in the Ottawa analysis.

Six O-D demand rates were used in the traffic assignment analysis to identify any demand-related trends. The percentage of trips for each O-D pair was kept constant, and only the total number of trips was varied. Table 1 lists the scale factors and the corresponding weighted average volume-to-capacity (v/c) ratio on the network for a UE traffic assignment based on travel time. The weighted average v/c ratio is the average volume-to-capacity ratio on all links weighted by the number of vehicles on the link. The first demand rate, with a scale factor of 1, represents a lightly loaded network, as illustrated by an average weighted v/c ratio of 0.19. When the scale factor is increased to 6, the average v/c ratio increases to 0.91.

TABLE 1 Key to Demand Rates

Demand Rate	Average Link
1	0.19
2	0.33
3	0.49
4	0.62
5	0.78
6	0.91

Traffic Assignment Based on Environmental Objectives on Large Networks

Common assignment techniques such as the Frank-Wolfe algorithm and the method of successive averages algorithm lend themselves well to traffic assignment based on environmental objectives. As discussed above all that is required is a link cost function based on CO production instead of one based on travel time.

Traffic assignments based on UE (travel time), SO (travel time), UE (CO production), and SO (CO production) were performed by using the ASSIGN traffic assignment model (7) on the Ottawa network. Both the Frank-Wolfe algorithm and the method of successive averages algorithm were used, and both gave approximately equivalent results. The solutions presented in this paper were derived by using the former algorithm. In every case the traffic assignment results based on environmental objectives and obtained by using Equation 2 as the link cost function met the underlying objective to minimize CO production. In the case of UE-CO the levels of CO production on all of the used routes for a given O-D were equal, and there were no routes that had lower levels of CO production. For the SO-CO examples the marginal CO production rates on all used routes were equal, and no unused route had lower marginal CO

production rates. In addition the total amount of CO produced decreased with each iteration of the algorithm and converged toward one value.

However some theoretical problems associated with using Equation 2 should be pointed out. It may be seen in Figure 3 that on Route 1 (the highway) the CO emissions per vehicle decrease as volume increases, reach a minimum, and then increase after that. This pattern is typical of pollution models in general because pollution levels tend to be highly correlated with fuel consumption. Fuel consumption is typically modeled as a function of speed, with some minimum rate occurring at an optimal speed that is typically in the range of 45 to 55 mph. For speeds on either side of the optimal speed fuel consumption increases at an increasing rate. As the volume on the link increases the speed decreases. This results in lower fuel consumption and hence lower CO emissions. Eventually the average speed decreases past the "optimal" velocity and the CO emissions begin to rise. This is illustrated by the convex shape of the CO production function for Route 1 in Figure 3. Note that the arterial link does not follow this pattern but rather increases with all volumes. This is because for this link the vehicles always travel below the "optimal" velocity because of speed limit constraints.

It is a well-known fact that the generalized cost functions used in UE and SO traffic assignments must be positive and must increase with volume (8). When they do not there is no guarantee that the resulting UE or SO program will have a unique minimum point. There are two important points relating to this last statement. The first is that the solution found by using the Frank-Wolfe algorithm may not be a global minimum but rather only a local minimum. The second point is that there is no guarantee that the link flows that are identified are unique. From a SO-CO perspective this is not overly critical in that in this paper the authors are only interested in system CO production and total travel time. Therefore at worst the SO-CO solutions are a conservative estimate, and theoretically there could be a "better" solution. However from a UE-CO perspective the results could be more subtle. Theoretically there could be an alternative set of link flows that results in the same or a smaller value of the objective function but that has lower (or higher) aggregate CO values than those reported in this paper.

It should be pointed out that when the conditions for a unique minimum (link cost function increasing and positive) are violated it does not automatically indicate that the resulting solution will not be a global minimum or unique. This is especially true for the examples used in this paper, in which the cost function decreases only marginally with travel time before increasing once the "optimal" velocity is reached. The fact that all of the environmental assignment solutions met the underlying objectives, as discussed earlier, was taken as an internal consistency check that the results presented in the following section are acceptable.

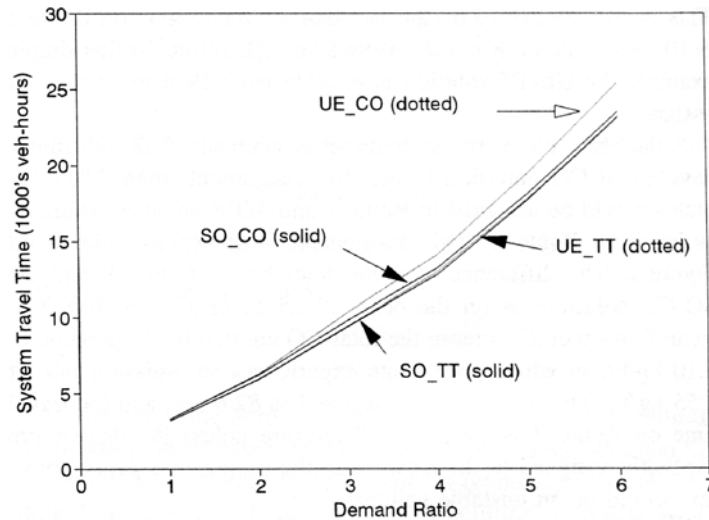


FIGURE 4 System travel time versus O-D demand ratio.

Ottawa Network Results

Figure 4 shows the system travel time results for each of the four traffic assignment scenarios as a function of the O-D demand ratio used. A number of trends are present that would be expected given past research in the area and general knowledge of macroscopic traffic assignment models.

These include the following:

1. As the demand increases the total system travel time also increases and at a slightly increasing rate for all four assignment scenarios.
2. The total system travel time is lowest for the SO assignment based on travel time.
3. The UE solution based on travel time is very similar to that for the SO assignment based on travel time.

As in the sample problem, the UE-CO solution produced similar (although slightly higher) total system travel time results compared with those produced by the UE-TT and SO-TT solutions. In addition, the SO-CO assignment produced the worst results with respect to total travel time on the system.

Figure 5 was created to better illustrate the differences between the various scenarios. Figure 5 illustrates the percent increase in system travel time for each traffic assignment scenario compared with the system travel time for the SO assignment based on travel time (i.e., the assignment with the lowest system travel time for a given demand level). The latter traffic assignment, SO-T-r, is therefore represented in Figure 5 by a straight-line function that is equal to zero. It should also be noted that although the travel time differences are not huge (i.e., <10 percent) the maximum amount of system travel time savings that may be achieved through IVHSs is usually considered to be in the range of 5 to 10 percent (9).

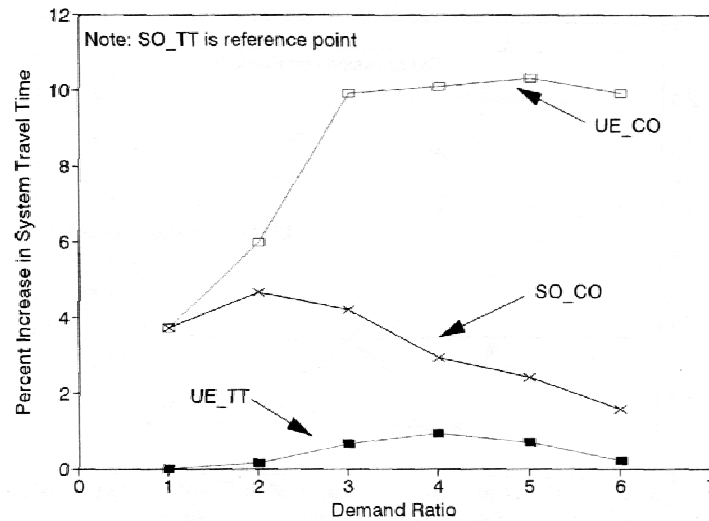


FIGURE 5 Percent increase in system travel time versus O-D demand ratio.

It may be seen in Figure 5 that the difference in the SO-TT and UE-TT solutions are minimal. However the relationship is concave, with the biggest difference, on the order of 1 percent, occurring at the middle demand levels. This may be explained by the fact that at low and high demand levels there is not as much route choice for the vehicles because the link travel times are not affected by their individual decisions. For example at low O-D demand levels every vehicle is basically assigned to the minimum path route that, because of the low volumes, does not Experience an appreciable rise in travel time. At high demand levels the results are more subtle. The link flows on the highways based on the UE-TT assignment are on average about 5 percent higher than those based on the SO-TT assignment. However because of the nature of the travel time function this difference in link flows results in only a marginal difference in aggregate travel time.

When the traffic is assigned to the network with the objective of an SO-CO assignment the total system travel time is on the order of 1.5 to 4 percent higher than that observed for an SO-TT assignment. In general as the demand rate increases this difference decreases. The link flows for the SO-CO assignment tended to be more similar to those for the UE-TT assignment, in which more vehicles are assigned to the highways than by the SO-TT assignment. This makes intuitive sense in that the link pollution-volume curve is relatively flat for highways, and therefore adding volume to highways does not appreciably raise pollution levels as much as adding vehicles to arterial roadways. On the basis of the results from Figure 5 it may be seen that for congested networks if policies were implemented to reduce overall CO emissions the actual amount of travel time in the system would increase on the order of 2 percent.

If the drivers were allowed to choose their routes individually and their decisions were based solely on the amount of CO produced (EU-CO), then the total system travel time would increase by approximately 4 to 10 percent above that if the drivers considered only the effect of travel

time. The trend generally increases with demand, although there appears to be a relative leveling off at a 10 percent increase once the O-D demand ratio reaches 3. Similar to the SO-TT assignment the UE-CO assignment also favors on average the use of arterial roadways. However the link flows are very dissimilar between the two assignments, and this results from the different generalized cost curves used and the underlying objectives in each assignment. It is this difference in link flows that is illustrated by the 10 percent difference in aggregate travel times

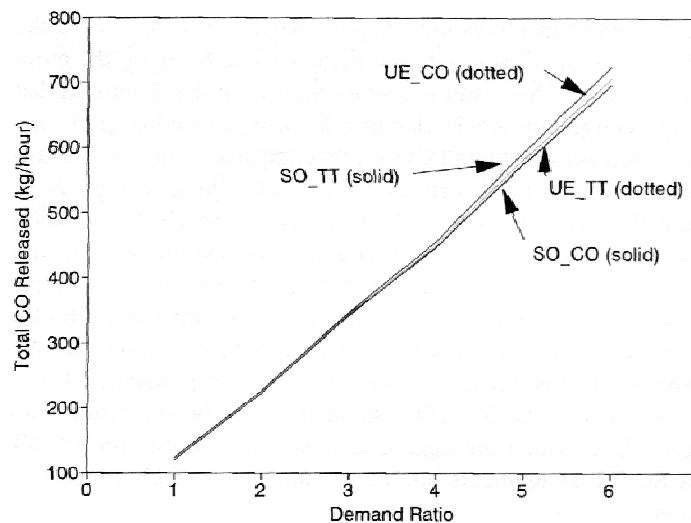


FIGURE 6 CO production versus O-D demand ratio.

for the UE-CO and SO-CO conditions.

Figure 6 illustrates the relationship between the total amount of CO released into the network for each of the four scenarios tested as a function of the O-D demand. It may be seen that the amount of CO released increases in a linear manner as the O-D rate increases. Figure 7 shows the percent increase in total CO produced for each scenario compared with that produced from the SO traffic assignment based on CO considerations (the lowest for a given demand level). The latter traffic assignment is therefore represented by a straight-line function that is equal to zero.

It may be seen that for high demand levels the UE-TT assignment results in CO output very similar to that from the SO assignment, which seeks to minimize this value (SO-CO). However at lower demand levels the difference in CO output can be on the order of 2.5 percent. This implies that in congested networks the UE-TT objective (i.e., what currently occurs) results in CO production levels that are approximately equivalent to what could be achieved if vehicles were directly routed through the network on the basis of minimizing CO production. As stated previously this is due to the similarities in link flows, in which in both cases the highways generally experienced higher volumes.

The opposite trend occurs for the UE-CO assignment when the drivers are taxed on the amount of CO that their vehicles produce and the drivers choose their routes solely to minimize this cost. At low demand levels the difference in CO emissions is minimal.

The difference increases in an approximately linear manner as the O-D demand increases, with the highest value being on the order of 1.5 percent. This result is counterintuitive in that it implies that at high congestion levels charging for CO production might actually increase the overall CO levels compared with those resulting from leaving the system as it is (all other things being equal). The difference is caused by the fact that for the UE-CO solution vehicles tended to utilize the arterial roadways more than they did for the SO-CO solution. At higher demand levels vehicles on arterial roadways tend to have higher pollutant emissions (all else being equal). A similar result was found in the two-node example analysis. The divergence between the total CO production levels for the UE-CO and SO-CO assignments may be contrasted with Figure 7, in which the aggregate travel time results for UE-TT and SO-TT assignments tended to converge at high congestion levels.

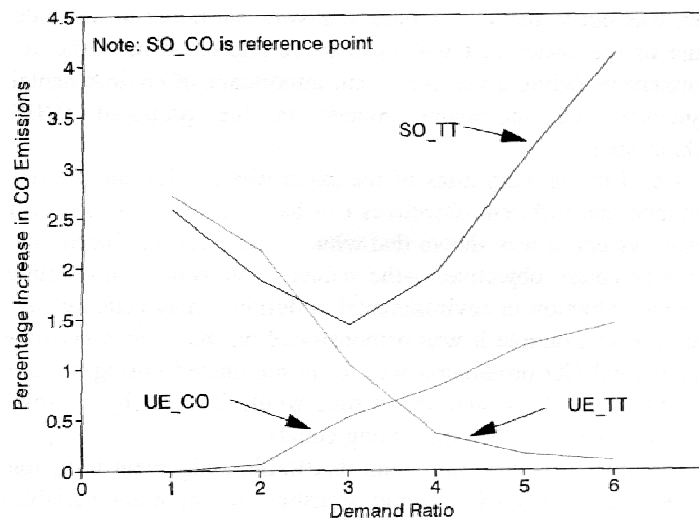


FIGURE 7 Percent increase in CO production versus O-D demand ratio.

In terms of total CO production the SO-TT assignment has the largest difference compared with the SO-CO assignment. The relationship is convex, with the difference first decreasing as demand increases until the demand ratio is 3 and increasing after that. The minimum difference is 1.5 percent, and the largest is approximately 4 percent, which occurs at the demand ratio of 6. Again this pattern is a result of the fact that in the SO-TT assignment the arterial roadways tend to have higher link volumes than the SO-CO assignment at high congestion levels. The results shown in Figure 7 imply that if an RGS was instituted with the sole purpose of minimizing travel time, other important objectives, such as reducing CO levels, might actually become worse.

CONCLUDING REMARKS

It can often be dangerous to generalize the results of macroscopic traffic assignment models to actual traffic systems. This would be especially true in this analysis, in which many simplifying assumptions are made (i.e., 100 percent homogeneous driver and vehicle populations) and very

simple production functions are used (i.e., Bureau of Public Roads travel time function and TRANSYT CO emission function). Perhaps more importantly vehicle emissions were assumed to occur uniformly along the length of the link rather than to have higher concentrations at the intersections, where traffic often stops. The intent of this paper, however, was not to derive a definitive answer but rather to illustrate some of the issues that will need to be addressed in traffic assignment modeling given the recent importance of environmental objectives and increasing interest in the proposed IVHS technologies.

One of the primary aims of the paper was to illustrate the importance that different objectives can have on the operation of a traffic system. It was shown that what would intuitively seem like complementary objectives-the reduction in system travel time and the reduction in environmental pollution-may actually conflict. As an example it was demonstrated on the Ottawa network that if total CO production was to be minimized during periods of congestion the system travel time would increase by approximately 2 percent (all things being equal),

Therefore it is very important that traffic assignment programs produce comprehensive output statistics of important variables such as travel time, pollutant emissions, and noise levels to enable transportation professionals to study any trade-offs that may be required. Related to this task is the fact that appropriate production functions need to be identified for these negative externalities that relate the important link attributes (travel time, stop time) to the amounts of pollutants produced. Traffic engineers need to adopt a more sophisticated generalized cost function that has a wide range of important parameters as opposed to one that is based solely on travel time. This may not be as straightforward as it appears because not only will the pollutants have different weights but these weights may be a function of the amounts of pollutants produced. For example a doubling of CO levels on a link may result in a quadrupling of importance of the CO level in the generalized cost function because of a nonlinear relationship between CO level and general health. In addition the generalized cost function may not strictly increase as a function of flow when pollution costs are involved. This could result in the need to develop assignment techniques different from those that have historically been used.

The second point that is raised in this paper is that there is a definite need to expand traffic assignment techniques to account for change-, in system objectives and changing technologies. To a certain extent this has been done (1-3,10) with respect to evaluating IVHS operations. However to date there does not appear to be any assignment models that assign traffic on the basis of an equity-as opposed to an equilibrium-objective function. As was demonstrated in the two-node network problem, there is a definite need to examine the effects that equitable objectives could have on traffic networks and whether there might be other methods of achieving the same results. For example it was demonstrated on the sample network that different techniques may be used to achieve the same goal. The CO emissions analysis showed that charging drivers for their individual production of CO and letting them make their own decisions (UE assignment) gave results equivalent to those of explicitly routing the drivers on the basis of considerations (SE assignment). Of course whether any of these patterns will hold for larger, more complex networks and for more realistic traffic assignment models needs to be studied.

ACKNOWLEDGMENT

Funding for this research has been provided by the Canadian Natural Science and Engineering Research Council.

REFERENCES

1. Rilett, L. R., and M. Van Aerde. Routing Based on Anticipated Travel Times. *International Conference on Applications of Advanced Technologies in Transportation Engineering*, Minneapolis, Minn., August 1991.
2. Mahmassani, H. S. Dynamic Models of Commuter Behavior: Experimental Investigation and Application to the Analysis of Planned Traffic Disruptions. *Transportation Research A*, Vol. 24A, No. 6, 1990.
3. Rilett, L. R. *Modeling of TravTek's Route Guidance Logic Using INTEGRATION Model*. Ph. D. thesis. Queen's University, 1992.
4. Rilett, L. R., B. G. Hutchinson, and R. C. G. Haas. Cost Allocation Implications of Flexible Pavement Deterioration Models. In *Transportation Research Record 1215*, TRB, National Research Council, Washington, D.C., 1989.
5. Penic, M. A., and J. Upchurch. TRANSYT-7F, Enhancement for Fuel Consumption, Pollution Emissions, and User Costs. In *Transportation Research Record 1360*, TRB, National Research Council, Washington, D.C., 1992.
6. Dowling, R., and A. Skabardonis. Improving Average Travel Speeds Estimated by Planning Models. In *Transportation Research Record 1366*, TRB, National Research Council, Washington, D.C., 1992.
7. Rilett, L. R., C. Benedek, and M. Van Aerde. *User's Guide to ASSIGN.- A Macroscopic Traffic Assignment Model*. Transportation Research Group, University of Alberta, 1992.
8. Sheffi, Y. *Urban Transportation Networks*. Prentice-Hall, Incorporated, Englewood Cliffs, N.J., 1985.
9. Ying, G. F., and T. M. Mast. Excess Travel: Causes, Extent, and Consequences. In *Transportation Research Record 1111*, TRB, National Research Council, Washington, D.C., 1987.
10. Chang, G., t. Junchaya, and L. Zhuang. In Integrated Route Assignment and Traffic Simulation System with a Massively parallel Computing Architecture. *Proc., Pacific Rim TransTech Conference*, Seattle, Wash., July 1993.

Publication of this paper sponsored by Committee on Transportation Supply Analysis.

APPLICATION OF DYNAMIC ASSIGNMENT IN WASHINGTON, D.C., METROPOLITAN AREA

E. DE ROMPH, H.J.M. VAN GROL, AND R.HAMERSLAG¹

A study in which the dynamic assignment model 3DAS was used as a planning tool is described. The Virginia part of the Washington, D.C.. metropolitan area was chosen for the study. This area offers a heavily congested urban network with several rerouting possibilities. On the basis of available data it was decided to calculate a morning peak hour from 5:00 until 11:00 a.m. in 24 periods of 15 min. each. The results show that the use of dynamic assignment for planning purposes can be very helpful. Dynamic assignment gives more detailed information than static assignment methods about the occurrences of traffic jams, and a more precise location and cause of congestion can be identified. Advanced traffic management system measures, introduced to alleviate the congestion, can be simulated, and all kinds of evaluations are possible, such as influences on travel time and jam length and effects of ramp metering and rerouting. Dynamic assignment, however, requires more accurate data and more computing time. Also very important is the ability to visualize the results. A dynamic assignment model gives flows in time. The best way to analyze the results is to present them in a movielike fashion. This requires a computer with a powerful graphics capability. For advanced traffic management systems to be successful more data and better (three-dimensional) origin-destination matrices are needed. New methods for origin-destination estimation and data from more induction loop, and probe vehicles will improve the reliability of the results.

This paper describes a study in which the dynamic assignment model 3DAS is used as a planning tool. The study has two objectives. The primary objective is to find answers to the following three questions:

1. Can dynamic assignment be used for planning purpose- ,?
2. Does dynamic assignment have an advantage above static assignment?
3. Is dynamic assignment a useful tool for investigating the effects of advanced traffic management systems (ATMSs)?

The secondary objective is to gain insight into the possibilities and problems associated with the application of 3DAS on large networks.

The model is applied to the southwestern part of the Washington. D.C., metropolitan area in the United States. This area was chosen because it offers a heavily congested urban network with several rerouting possibilities. Several ramp metering installations are in operation, and parts of the freeways are monitored. The data used for this research were obtained from the Virginia Department of Transportation (VDOT) and the Metropolitan Washington Council of

¹ E. de Romph and R. Hamerslag, Department of Infrastructure, Faculty of Civil Engineering, Delft University of Technology, P.O. box 5048, 2600 GA Delft, The Netherlands. H.J.M. van Grol, Department of Computational Physics, Faculty of Applied Physics, Delft University of Technology, Lorentzweg 1, 2628 CJ Delft, The Netherlands.

Governments (COG). A small portion of the study area is monitored by induction loops. One-minute data from these induction loops were used to derive the departure time functions and to validate the calculation results.

The research was conducted during a 4-month visit to the Center for Transportation Research at Virginia Polytechnic Institute and State University (Virginia Tech). In accordance with the objectives the study is meant only as an example of the use of dynamic assignment as a planning tool. Because of the lack of data and the short study time the calculated results are not suitable for use in making serious planning decisions. The results, however, do permit one to determine the usefulness of dynamic assignment for planning purposes.

Briefly discussed are the 3DAS model, the research approach, and how the data were derived. Apart from a static assignment, three different scenarios are calculated: a morning peak hour scenario, a scenario with several ramp metering installations, and a scenario with an incident. The results of the model for these scenarios are reported.

3DAS MODEL

The 3DAS model is based on the work carried out by Hamerslag and Opstal (1) and Hamerslag (2,3). The basic feature of a dynamic assignment model is the partitioning of time into small slices, usually referred to as *periods*. Over the last 2 years the model has been improved, in particular its dynamic aspects. The 3DAS model has been described by de Romph et al. (4,5) and by van Grol (6).

The model determines the flow distribution in the network with an iterative process. In each iteration the shortest paths in the network are calculated for all origin-destination (OD) pairs and for every departure period. The link parameters are defined separately for each period. The properties of the network and the travel demand are presumed to be given.

The basic iteration scheme in Figure 1 is essentially the same as that for static assignment models. The difference lies in the all-or-nothing-in-time module. In this module an extra iteration over the departure period is needed, and the shortest path must be found and the assignment must be performed in time.

The paths are defined by using the travel time on a link in the period in which the traffic actually traverses the link; that is, the trajectory that the traffic follows in time is calculated. The network is loaded on the basis of these trajectories. During the assignment the contribution of a traveler to the traffic load on a link in a certain period is determined by calculating the duration of the traveler's presence on that link in that period. If one focuses on one traveler, two situations occur:

1. Several links are covered in one period. In this case the traveler is present on the link for only a part of the period, and therefore should be assigned to the link for only this part of the period.
2. One link is covered in several periods. The traveler is present on the link during multiple periods and should be assigned to the entire link for each individual period.

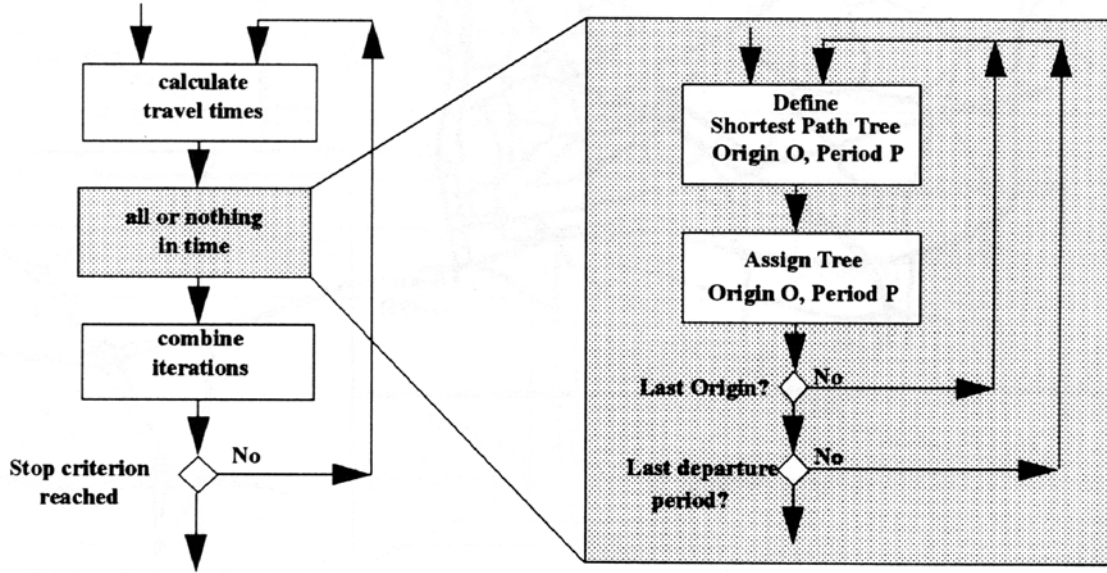
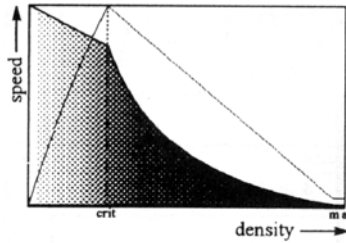


FIGURE 1 Iteration scheme.

At the start of each iteration the travel times on the links are derived from the load of the previous iteration. For each link, the travel time is calculated with a speed-density function. A relation between speed and density instead of the traditional relation between travel time and flow is used. This allows modeling of a decreasing flow in the case of congestion. The conservation of traffic and the continuity of flow are maintained. In case of overflow, the overflow is assigned to the preceding link on the path in the same period. The stop criterion is reached when there is no difference in the resulting flows between two successive iterations.

The 3DAS model has been tested on several small networks (4). Several parameters of the model were calibrated by using these networks. The initial settings of these parameters followed from these tests and were not changed for the study described here. A speed-density function of the following form is used:

$$v(\rho) = \begin{cases} v^{\max} \cdot \left(1 - \frac{\rho}{\rho^{\max}}\right) & 0 < \rho < \rho^{\text{crit}} \\ v^{\max} \rho^{\text{crit}} \cdot \left(\frac{1}{\rho} - \frac{1}{\rho^{\max}}\right) & \rho^{\text{crit}} < \rho < \rho^{\max} \end{cases}$$



where

v^{\max} = free-flow speed,
 ρ^{crit} = critical density, and
 ρ^{\max} = maximum density.

RESEARCH APPROACH

In accordance with the objectives of the study, the following research approach was set up. The first objective consists of the following three questions:

1. Can dynamic assignment be used for planning? Dynamic assignment can be used for planning if, given a network and a traffic demand, it can predict a correct distribution of traffic flow. Since for long-term purposes the traffic demand will represent the average demand, the expected traffic distribution will also be average. This is in contrast to real-time applications, when the results should be based on the actual situation at that moment. To validate the model the average traffic demand and a measured traffic distribution averaged over a longer period are required.
2. Does dynamic assignment have an advantage above static assignment? There are several (well-known) problems with static assignment models. A static assignment model
 - Can give wrong results when congestion occurs. Because traffic is assigned along the complete route, a car can contribute to more than one congestion at the same time.
 - Cannot correctly show the effects of a variable traffic demand.
 - Cannot correctly show the effects of temporal disturbances such as roadworks or accidents.
 - Cannot predict queue lengths and cannot show how a growing queue can limit the capacity of upstream junctions.

The authors determined whether dynamic assignment can solve these inconsistencies and how it will improve the decision making for planning.

3. Is dynamic assignment a useful tool for investigating the effects of ATMSS? The model has been extended to model several ATMS instruments, such as ramp metering, rerouting, and tidal flow. To answer the question two tests were executed. The first scenario considered several ramp metering installations, and the second scenario considered an accident at one of the freeways. For the second scenario the effects of diversion measures are reported.

Since the network used for the study is fairly large the secondary objective of this research, to gain insight into the possibilities of and the problems associated with dynamic assignment applied on larger networks, is also satisfied by the research approach described above.

DATA

The study area covered the eastern part (Virginia part) of the Capital Beltway around Washington, D.C. The major Interstates are I-95, I-395, I-66, and I-495; a large part of the arterial network was also included.

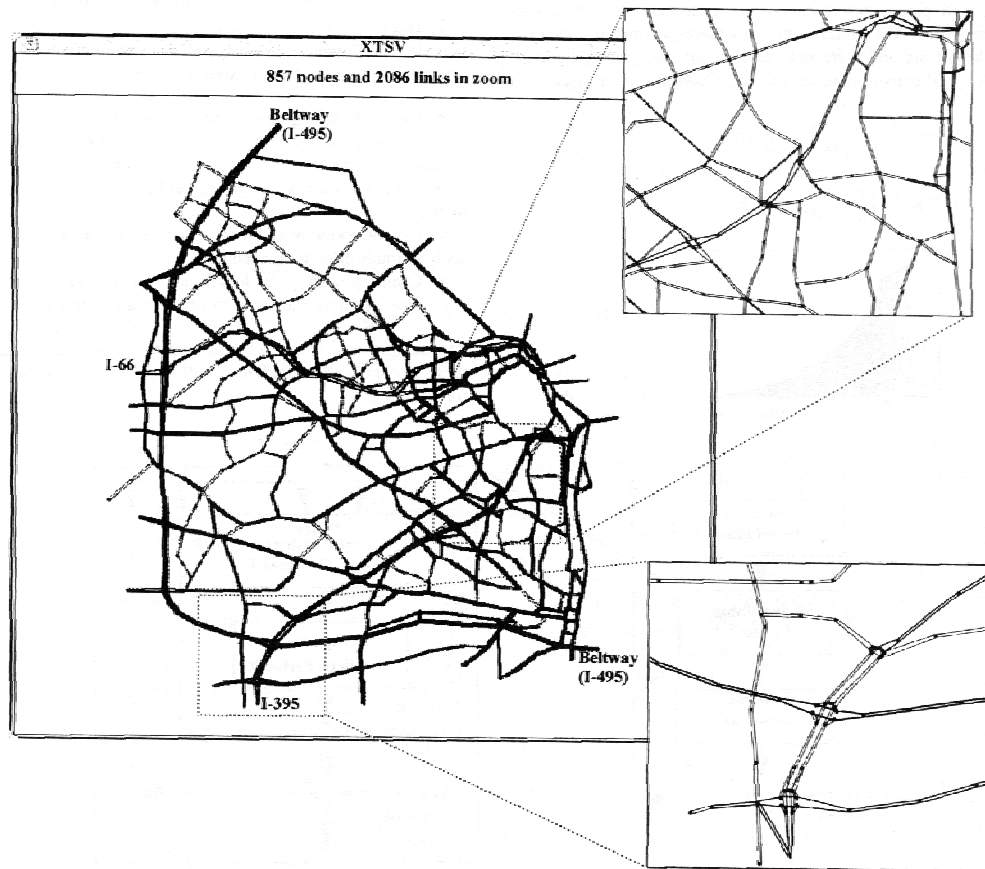


FIGURE 2 Study network.

Network

Figure 2 represents the network used for the study. The network consists of 857 nodes and 2,086 links. There are 180 zones. Most freeway intersections are represented in a fairly detailed way. Examples of two of these intersections and their detailed representatives are given in Figure 2. Each line in Figure 2 shows a separate one-directional road consisting of from one to four lanes.

The 2,086 links are divided into 13 types, each representing a certain road type. AU of the links in one type have the same attributes. The attribute for the capacity is not given but is derived from the maximum density, the maximum speed, and the speed-density function.

OD Matirix

The network is not accompanied by a matching dynamic OD matrix. This OD matrix must be constructed from other data sources.

The best OD matrix available was a (static) 24-hr matrix that covers a much larger area. This OD matrix resulted from a study by COG. The OD matrix for the study area had to be extracted from this OD matrix. To make the OD matrix dynamic, departure time functions were used. A departure time function describes for each OD pair the portions of the amount of traffic departing in each period. These departure time functions can be estimated and calibrated with link volume data.

The COG study (7) was done with 1990 as the base year and comprised 293 districts (1,478 zones), which covers the entire area of Washington, D.C., and several surrounding jurisdictions in Virginia and Maryland. The network covered 5,983 nodes and 18,104 links.

The model used by COG for the trip generation, distribution, and mode choice was a gravity model and was calculated at the district level. The districts were then split into zones via land use factors. For production these land use factors were based on household and groups-quarter population. For attraction they were based on office, retail, industrial, and other employment. The resulting OD matrix had 1,478 zones.

The network used for the study (Figure 2) is only a part of the COG network, so the OD matrix for the smaller network (180 zones) had to be derived from the large OD matrix (1,478 zones). All trips made within the study network are easily derived. All trips entering, leaving, or passing through the study network were derived by a selected link analysis. To perform the selected link analysis the OD matrix is assigned to the network with a static all-or-nothing assignment. The shortest path is found by using the actual speeds in the network. These actual speeds were derived from the static assignment done by COG. For all OD pairs crossing the selected links the origin and the destination are stored. The entering and exiting links become new origins and destinations, and the trips are summed. By using this method all entering and exiting traffic is aggregated to the links in which it exits or enters the subnetwork.

Derivation of an OD matrix for the subnetwork by this method has one major drawback. Because an all-or-nothing assignment is used no alternative routes are chosen for OD pairs. To minimize the effects of this problem some links are added to the subnetwork to allow a diversion for some origins to different links to enter the network.

Induction Loop Data

The Northern Virginia Traffic Control Center controls a part of the freeway system in northern Virginia. The freeways covered are I-66 and I-395. These freeways are equipped with several hundred induction loops. One minute of data for fixed portion of these induction loops can be downloaded on a data tape. Unfortunately, the Traffic Control Center is not yet fully equipped, and the downloading of data from induction loops is therefore not easy. Only one tape (1-day) was available for the present research. Although the traffic patterns of this 1 day were not sufficient to derive any statistical information, they were the best data available. The tape used for the study contained data measured on Monday, December 7, 1992, from 4:00 p.m. until 1

1:00 a.m. the next day. The number of vehicles that passed was registered and downloaded every minute.

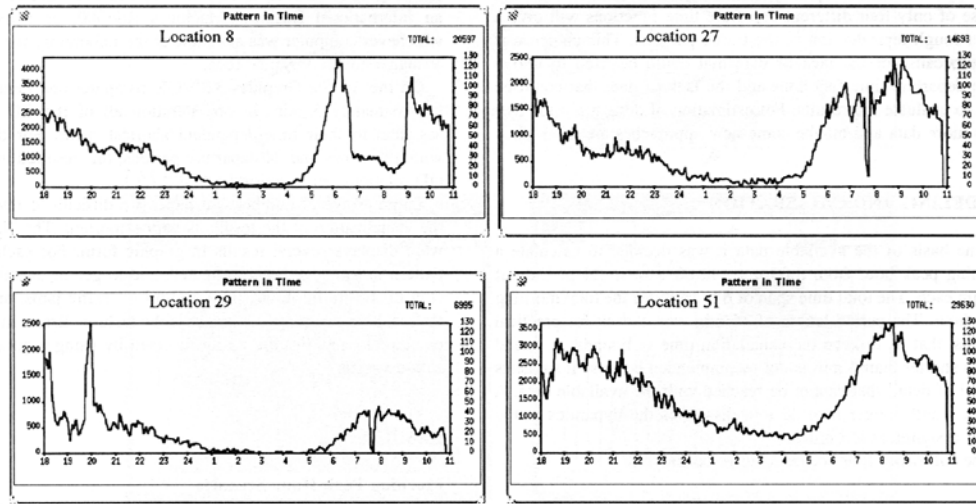


FIGURE 3 Traffic flow from 4:00 p.m. to 11:00 a.m. (next day) for locations 8 and 27 on I-66 eastbound and locations 29 and 51 on I-66 westbound.

Figure 3 gives an impression of the traffic patterns at several locations on I-66. The x-axis shows the time in hours. The registration started at 4:00 p.m. and lasted until 11:00 a.m. the next day. The y-axis shows the flow in vehicles per hour. For each direction two graphs are shown. The first graph is located at the beginning of the freeway, and the second graph is located near the end of the freeway. The locations of the induction loops are displayed in Figure 4.

Figure 3 shows that the peak hour starts at ± 5.00 a.m. At location 8 the flow increases in approximately 1 hr to 4,500 vehicles/hr. At $\pm 6:00$ a.m. some kind of congestion occurs and the flow drops rapidly (possibly an incident). After $\pm 9:00$ a.m. the flow increases again. The end of the peak hour is at approximately 11:00 a.m. At location 27, which is farther downstream I-66, the flow increases to $\pm 2,500$ vehicles/hr. The two graphs for locations 29 and 51 shown that the flow on I-66 westbound is lower and that no congestion occurs in this direction. At location 29 the flow increases to $\pm 2,000$ vehicles/hr. At location 51 the flow increases to $\pm 3,500$ vehicles/hr.

On the basis of the induction loop data it was decided to calculate a morning peak hour from 5:00 to 11 a.m. This time period captures the total morning peak hour, and the graphs show that before 5:00 a.m. the network is still reasonably empty. This has the advantage that the calculations can be started with an empty network.

Departure Time Functions

To use a static OD matrix as a substitute for a dynamic OD matrix departure time functions are required. A departure time function is a discrete function that determines for each period the percentage of the OD value that departs during that period. To derive these departure time functions induction loop data can be used.

One departure time function for all OD pairs will not give a realistic representation of the dynamic OD matrix for the peak hour. The departure time function, of individual OD pairs can be quite different. Figure 3 shows that the volume of traffic on I-66 traveling westward is lower in the morning peak hour and higher in the evening peak hour and that traffic departs according to a different departure time function. The same observation was made for I-395. This requires at least different departure time functions for traffic entering Washington and traffic leaving Washington. For this reason the OD matrix is split into four major trip types. For each type a different departure time function is used.

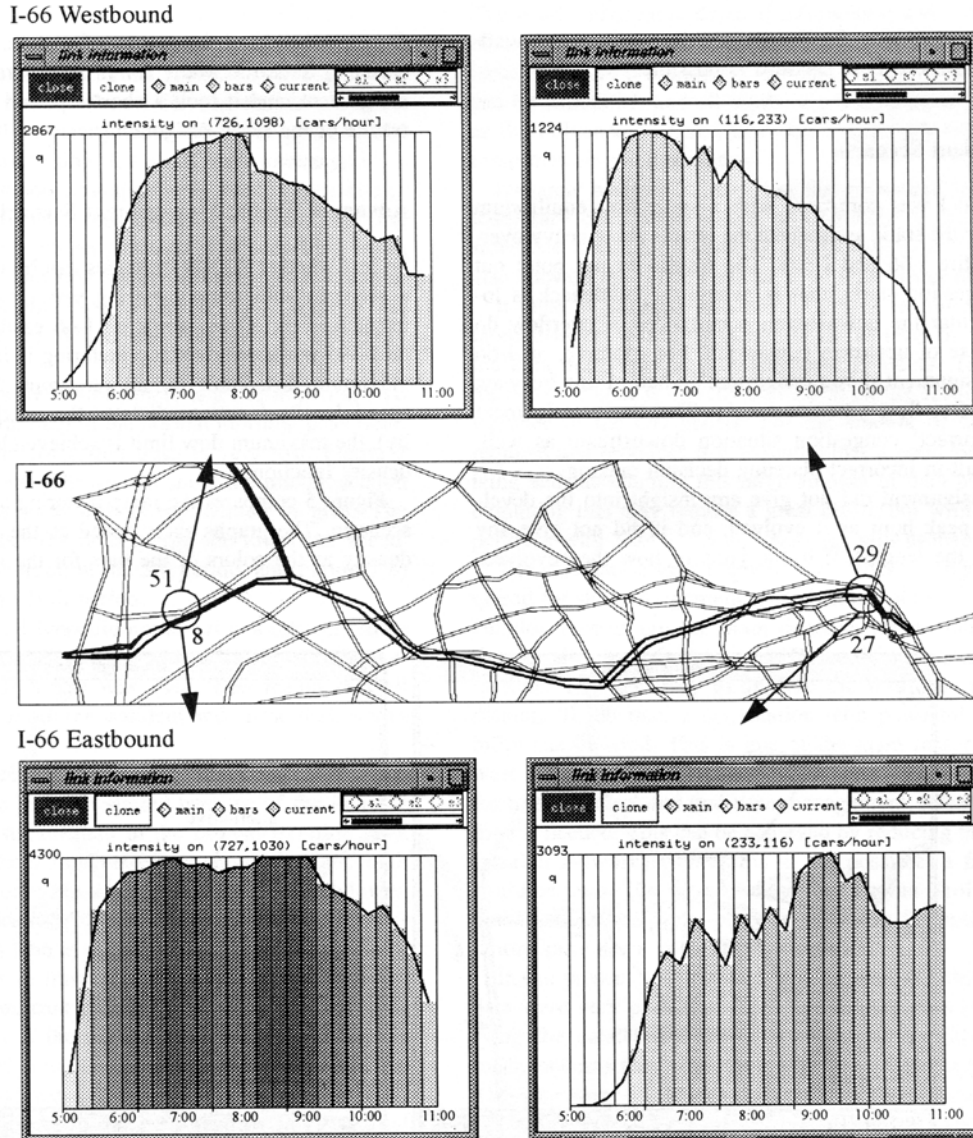


FIGURE 4 Flow calculated by 3DAS for peak hour from 5:00 to 11:00 a.m. at four locations (51, 8, 29, 27) on I-66.

Use of only four different departure time functions will give a rather rough reproduction of the traffic patterns. This choice was made because of the lack of data that could be used to derive more departure time functions and the lack of data that could be used to evaluate the results. For estimation of departure time patterns more data and maybe some new approaches are desirable.

MODELING AND CALIBRATION

On the basis of the available data it was decided to calculate a morning peak hour from 5:00 until 11:00 a.m. in 24 periods of 15 min. each. The total time span of 6 hr captures the total morning peak hour. The period length of 15 min. was chosen for practical reasons, that is, to keep the calculation time in bounds. A period length shorter than 5 min. is not recommended because it suggests a level of detail that cannot be reached with the available data. A period length longer than 20 min. dissipates the dynamics in the traffic assignment too much.

The following four scenarios were considered:

1. The first scenario is meant to achieve a reasonable reproduction of the morning peak hour. The departure time functions are calibrated with induction loop data, and the resulting flows are compared with the induction loop data. By adapting the departure time functions it is possible to reproduce the induction loop data at the beginning of a route. When the flow pattern farther downstream on that route still matches the induction loop data, this is considered a good result. The flow pattern can be tested at several locations on the following time-varying form, the average height of the flow, and the moments of sudden changes in the flow. Since only day of induction loop data was available and no information about weather or incidents was available, these data do not represent an average flow pattern. Only a rough re-production of volume patterns can be expected.
2. The second scenario is a static equilibrium assignment comparing the results with those of a dynamic assignment. The advantages and disadvantages of time variation are studied.
3. The third scenario introduces ramp metering at all ramps on I-66 eastbound and at all ramps on I-395 northbound. The influences on queue length, travel time, and diversion behavior are investigated.
4. The fourth scenario introduces an accident at one of the freeways (I-66). For this scenario two different situations are calculated. In the first situation the drivers are unaware of the accident. This is simulated by using initial travel times for the section with the accident. In the second situation the drivers are assumed to be fully informed. Here an equilibrium assignment is used.

The third and the fourth scenarios investigate the possibilities of dynamic assignment for ATMSs. The input data used for these scenarios are the same as those used for the morning peak hour scenario (scenario 1). The departure time functions and the OD matrix remain unchanged.

HARDWARE AND SOFTWARE

The model is implemented as an X-window program for the UNIX operating system. Several different computers were used to run the program. We used a Silicon Graphics 32OVGX or

INDIGO, an International Business Machines RS6000, or SUN Sparc2, whichever computer was available at the Laboratory for Scientific Visualization at Virginia Tech.

On the Silicon Graphics 32OVGX computer one iteration took approximately 5 min. In one iteration all of the OD pairs are assigned to their time-dependent shortest paths. For this study with 180 zones and 24 departure periods, this resulted in 692,040 OD relations per iteration.

Large arrays of numbers on paper are difficult to interpret, so the visualization of the results is very important. The 3DAS software displays several results in graphic form. For each link the pattern in time can be investigated, and to get an overall impression of the traffic flows, the build-up of traffic jams, and so on, the results are displayed in a movielike fashion. Errors in the input or other anomalies are easily detected by using a good visualization system.

RESULTS

Morning Peak Hour Scenario

On the basis of the OD matrix, the departure time functions, and the network attributes a dynamic assignment was done. Heavy congestion was found on I-66 and I-395 going into Washington, D.C.; low levels of congestion were found at several locations on the beltway and on certain arterials. The movielike representation showed quite clearly where congestion started and how it evolved. To give an impression of the results, the flow patterns at two locations along I-66 (Figure 4) are shown.

Figure 4 shows the flow (intensity) at four different locations on I-66. The x-axis represents the time, and the y-axis shows the flow. Each bar represents a time period of 15 min. The heights of the bars measure the flow, whereas the colors of the bars show the density. Light grey represents a low density, and dark grey represents a high density. By using this representation the difference between a low flow caused by a high density (dark grey) and a low flow caused by a low density (light grey) can be discriminated.

Figure 4 represents the same locations on I-66 as the induction loop graphs in Figure 3.

Comparing the graphs in Figure 4 with the induction loop graphs in Figure 3, a reasonable reproduction of the traffic distribution was found to be possible. On I-66 eastbound, however, the induction loop data show heavy congestion with a low flow (almost zero). On the basis of the low flow downstream, one may assume that there was probably some kind of incident during that day. In the simulation a higher flow downstream was found. If there really was an incident the differences between the model and the induction loop data are explainable. To validate the result the flow pattern on the freeways were compared with the induction loop data at several places along I-66 and I-395. In general a fairly good match at I-66 and I-395 was achieved.

The speed results for the normal peak hour scenario are displayed as a solid line in the same graph. The x-axis represents time.

Figure 5 shows a location halfway on I-395 and one downstream on I-395. The two graphs demonstrate that there was a noticeable impact. Both locations show slight improvements in speed. In Figure 5(a) the temporal decrease in speed at 8:00 a.m. in the normal peak hour (solid

line) is no longer there. At the other location [Figure 5(b)] there is an improvement in speed almost over the total duration.

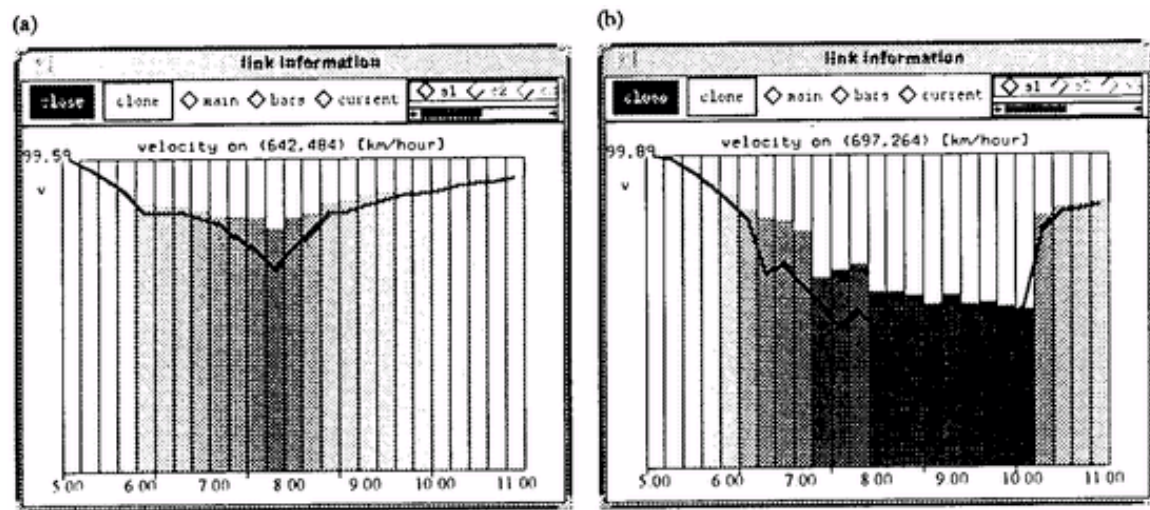


FIGURE 5 Velocity (km/hr) calculated by 3DAS at two locations on I-395. In bars, the ramp metering scenario is displayed. The solid line displays the velocity during normal peak hour.

Figure 6 shows the impact that ramp metering has on the arterial network. Figure 6 displays a location at the end of I-395.

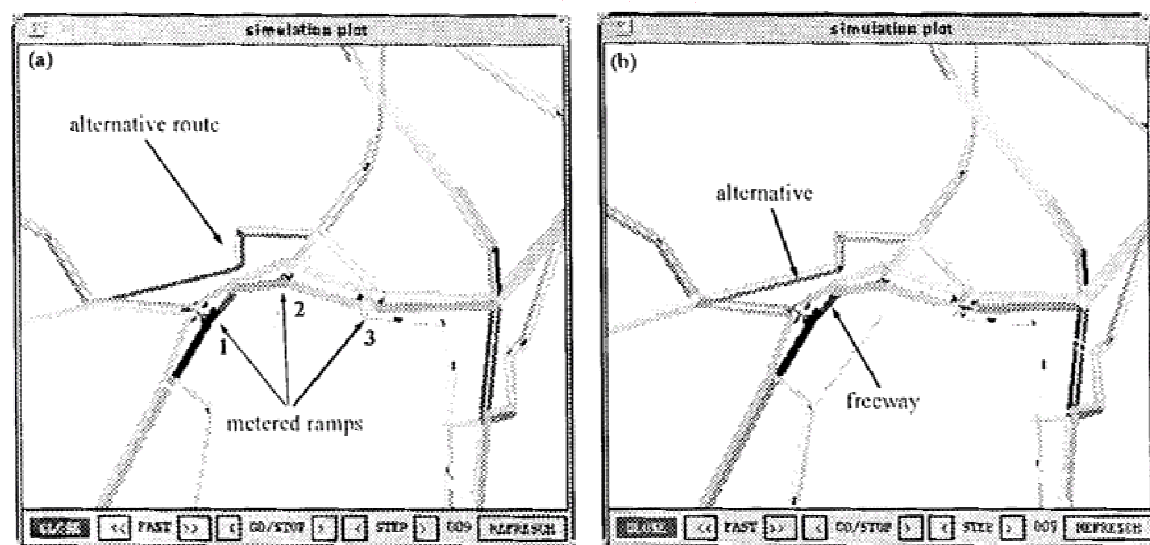


FIGURE 6 Rerouting behavior. Density in 9th period (7:00 to 7:15): (a) ramp metering scenario; (b) normal peak hour scenario.

Figure 6 was chosen to illustrate that because of ramp-metering alternative routes parallel to the freeway could be chosen. Figure 6(a) shows a slightly darker grey (higher density) than Figure 6(b) on the alternative route. On the freeway a slightly lower flow is detected. The values show that traffic is avoiding Ramp 1 and that a higher density is found on the alternative route.

Accident Scenario

To test whether the effects of incidents can be investigated with 3DAS, an accident was simulated on I-66. The accident was introduced by decreasing the capacity for a link by 60 percent. The OD matrix and the departure time functions were unchanged.

Two different route choice strategies were used. One strategy used the same routes that were chosen during a normal morning peak hour; the other route choice strategy was according to an equilibrium assignment. The first scenario represents a situation in which the accident is unknown to the travelers, whereas the second scenario is one in which each traveler is optimally diverted.

In the first scenario (no diversion) there is a traffic jam at I-66 that grows farther upstream than in the normal morning peak hour. The average speed of the congested links is very low. Figure 7 shows the situation on I-66. The graphs show the middle section of I-66. The density for each link is represented in grey. The darker the grey, the higher the density and the lower the speed. Figure 7(a) shows the situation in the 5th period, and Figure 7(b) shows the situation in the 10th period.

In the first scenario the drivers did not divert to a different route because they were not aware of the accident. In the second scenario an equilibrium assignment was used. This means that all travelers were informed about the accident and chose their routes accordingly.

The equilibrium assignment gave some remarkable results. The total length of the traffic jam that started because of the incident did not grow farther upstream than in the normal morning peak hour. Comparison with the normal peak hour shows that the length of the queue is in fact shorter but the average speed is much lower. Arterials around the location of the accident all have heavier loads. Figure 7(c) shows I-66 at the 10th period for this scenario.

When the travel times to traverse the entire length of I-66 are compared there is a significant difference between the two accident scenarios. In Figure 8 the normal peak hour travel time is compared with the travel times in the accident scenario and the accident with diversion scenario.

The free-flow travel time on I-66 is 11.5 min. For the normal morning peak hour it takes approximately 18 min. to traverse I-66 for traffic that departs at 7:30 a.m. In the case of the accident, when the traffic is rerouted, the travel time increases significantly, although the total length of time of the traffic jam is the same. When the traffic is not rerouted the travel time to traverse I-66 increases to almost an hour for traffic that departs at 7:45 a.m.

Figure 8 shows that the travel time is shortest during a normal peak hour. The scenario with the accident gives a travel time approximately three times as long. When diversion is allowed the travel times are approximately 1.5 times as long. This case shows an improvement of travel time by approximately 50 percent. Of course, this is an extreme case. The worse case is compared

with the optimal one, and there seems to be enough capacity on the arterial network, which could not be validated.

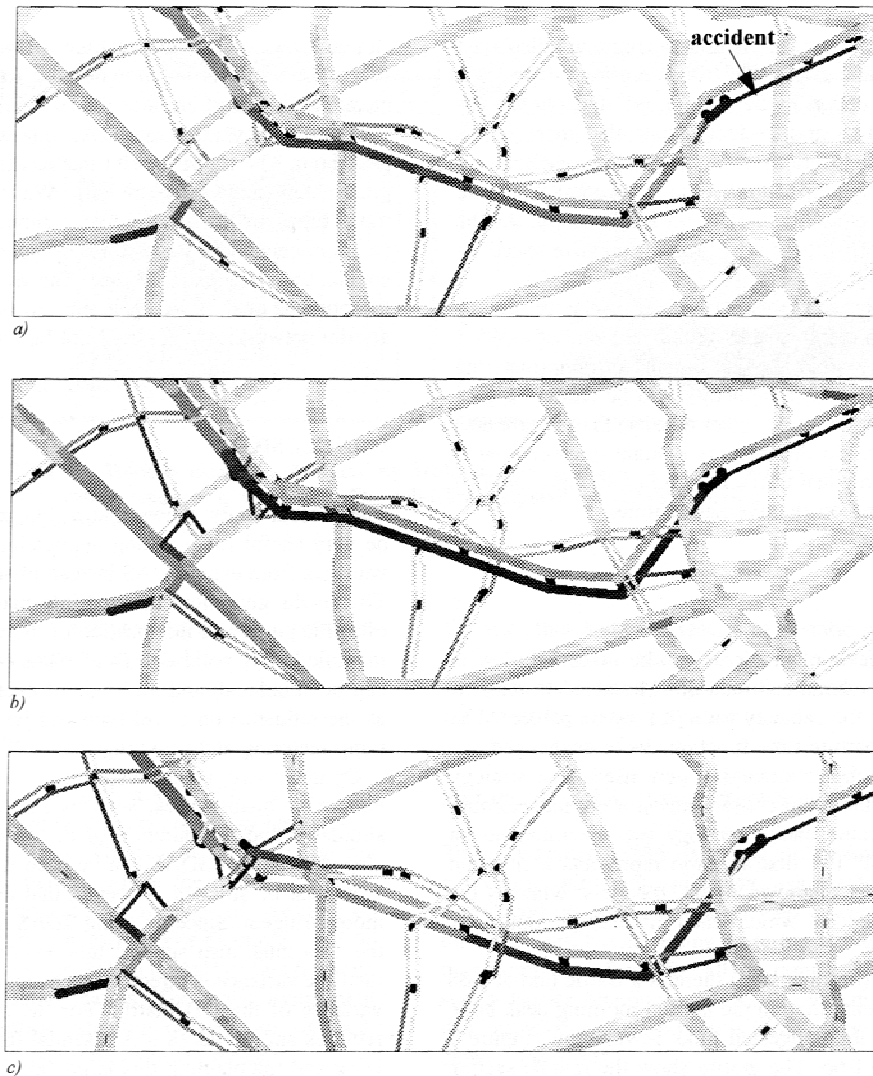


FIGURE 7 Accident at I-66: (a) 5th period, 6:15 a.m., no diversion; (b) 10th period, 7:30 a.m., no diversion; (c) 10th period, 7:30 a.m., with diversion.

CONCLUSION

The present study shows that a dynamic assignment model can be very useful for planning applications. A number of clear advantages from using 3DAS instead of static assignment are given. The results give more detailed information about the occurrences of traffic jams, and the location or the cause of congestion can be identified more precisely. To alleviate congestion ATMS measures can be simulated, and all kinds of evaluations are possible, such as the influence on travel time and jam length and the effects of ramp metering and rerouting.

Dynamic assignment also has the advantage that all kinds of temporary disturbances, such as accidents or roadwork, can be simulated and the duration of delays can be derived. The study also showed that 3DAS can be used with larger networks.

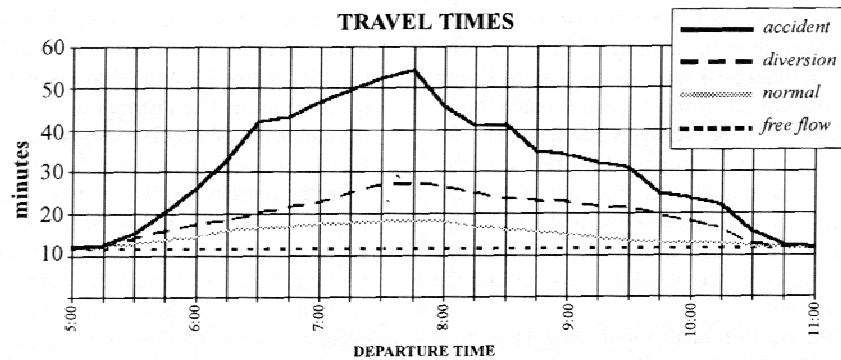


FIGURE 8 Travel times to traverse I-66 for four scenarios: free flow situation, normal peak hour, accident scenario, and accident with diversion scenario.

It must be stressed, however, that data requirements are much more stringent. Since by using 3DAS the level of detail is higher, the data must also support this level of detail. The accuracy of the time variance is directly dependent on the accuracy of the time variance of the OD matrix. For the amount of data that 3DAS requires and produces it is essential that a good system of organizing and maintaining this large amount of data be found. In the beginning this may require a great effort, but with increasing experience with 3DAS this disadvantage will probably disappear.

The calculation time required for 3DAS is longer than that required for static assignment. For planning purposes, however, calculation time is not the main concern. Much more important is the visualization of the results. Dynamic assignment gives flows in time. The best way to analyze the results is in a movielike fashion. To do that, a workstation with powerful graphics capability can be used. This is one of the main reasons workstations were used for the research described here. When the model is used for traffic control and real-time management, a faster computation time is needed. This can be achieved by reducing the problem size (smaller network). When this is not possible, a faster computer could be used. The research described by van Grol (6) toward the development of special-purpose hardware for assignment calculations provides a cost-effective solution.

In the present specific study the amount and the quality of the data were very poor, and there were limited possibilities for verifying the data. Since the authors had no insight into the local traffic patterns they could not judge the quality of the OD matrix. The time spent on this research was too short to make any serious planning decisions. The study is therefore primarily meant to investigate the usefulness of dynamic assignment for planning purposes. For real planning decisions a more elaborate study is required.

For ATMSs to be successful there is a large demand for more data and better (three-dimensional) matrices. New methods for OD estimation and data from more induction loops and probe vehicles will provide better results in the future.

With more time, more knowledge of the local study area, and more induction loop data the model has the potential to provide reliable information for real planning strategies and driver information systems.

ACKNOWLEDGMENTS

The authors thank the Netherlands Organization for Scientific Research and the University-Funds Delft for their financial support in making the 4-month visit to Virginia Tech possible; Virginia Tech for allowing the authors to use all of the facilities at the university; the Center for Transportation Research at Virginia Tech for their scientific support and their hospitality; the Laboratory for Scientific Visualization at Virginia Tech for making their hardware available and for their enthusiasm for the visualization of traffic; VDOT and COG, who provided us with the necessary data; and Peter Maas (applied physics student), who worked hard to get all of the necessary software running.

REFERENCES

1. Hamerslag, R., and P. C. H. Opstal. *A Three-Dimensional Assignment Method in Timespace* Report 87-94. Faculty of Mathematics and Informatics, Delft University of Technology, Delft, The Netherlands, 1987.
2. Hamerslag, R. *Dynamic Assignment in the Three Dimensional Time-space, Mathematical Specification and Algorithm*. USA-Italy Seminar on Transportation Networks, Naples and Capri, Italy, 1989.
3. Hamerslag, R. Dynamic Assignment in the Three Dimensional Time-space. In *Transportation Research Record 1220*, TRB, National Research Council, Washington, D.C., 1989.
4. de Romph, E., H. J. M. van Grol, and R. Hamerslag. A Dynamic Assignment Model for Short Term Prediction. Paper presented at Capri Seminar on Urban Traffic Networks, July 5 to 8, 1992, Italy.
5. de Romph, E., H. J. M. van Grol, and R. Hamerslag. 3DAS-3 Dimensional Assignment-A Dynamic Assignment Model for Short Term Predictions. 39th North American Meeting of the Regional Science Association International, Chicago, Ill., November 1992.
6. van Grol, H. J. M. *Traffic Assignment Problems Solved by Special Purpose Hardware with Emphasis on Real Time Applications*. Ph.D. thesis. ISBN 90-9005441-3, 1992.
7. Updates to COG/TTB Travel Forecasting Procedures, Highway Applications, 1990 Validation and 2010 Simulation. Draft. Washington Metropolitan Council of Governments, March 1993.

Publication of this paper sponsored by Committee on Passenger Travel Demand Forecasting.

MULTIPERIOD NETWORK IMPROVEMENT MODEL

CHIEN-HUNG WEI AND PAUL SCHONFELD¹

As traffic demand increases over time, improvements to existing transportation networks must be considered for enhancing efficiency, capacity or both. Because of limited resources even justifiable projects may have to be implemented gradually. The selection and timing of improvement projects are very important to ensure the most cost-effective investment plan. Conducting this task for transportation network is particularly challenging since the project effects tend to be inherently interdependent. By inadequately estimating project impacts during intermediate periods most existing methods tend to generate inappropriate improvement plans. The study developed a multiperiod network design problem model for the dynamic investment problem. A branch-and-bound algorithm was designed to determine the best project combinations and schedules. An artificial neural network model was used for estimating multiperiod user costs. The proposed model can efficiently handle the interdependencies among projects and demand change in each period. This method can be used for programming various transportation network improvements or transformations.

Investing in transportation systems to accommodate the increasing demand over time is one of the major issues for public agencies. Because of resource and other physical limitations, selecting the optimal project combination and implementation timing is very important for such programs. This problem is particularly challenging since most network projects are highly interdependent.

Evaluating the interrelations among projects of interest is often a critical issue for investment decisions. Several researchers have sought to derive appropriate expressions for various interrelations among projects. However their efforts have not yielded significant breakthroughs. Neither of these results is satisfactory in transportation networks where projects tend to affect each other. The network effects that cause such interrelations cannot be examined by simple analytical models. Therefore the interdependent terms will not be estimated for any project combinations in this paper. Instead the differences between various aggregate effects will be computed and used for comparing the effectiveness of various project combinations.

Existing methods tend to ignore the intermediate period conditions and hence may lead to inappropriate solutions. Since traffic demand may not increase smoothly over time and throughout the entire network, we should consider the effects of demand changes on networkwide operations. Even when the demand increases smoothly the resulting network equilibrium could be significantly different in each period because of motorist route choice behavior and the changing set of projects already implemented. Hence explicit consideration of intermediate-period conditions is essential in economic evaluations.

¹ C.-H. Wei, Department of Transportation and Communication Management Science, National Cheng Kung University, Tainan 70101, Taiwan. P. Schonfeld, Department of Civil Engineering, University of Maryland, College Park, Md. 20742.

It may be realized through the above discussions that the most suitable timing of various improvements is really dependent on many factors. Therefore a model for the multiperiod network design problem (MPNDP) is developed for programming transportation network improvements. This model includes the desirable features of simultaneously determining the best combination of projects and schedules.

LITERATURE REVIEW

Although a number of project selection studies have been made it seems that relatively little effort has been devoted to assessing interactions among projects (1). Researchers typically deal with simple interrelations [e.g., see the papers by Fox et al. (2) and Gomes (3)] or assume that such information is exogenously provided [e.g., see the paper by Carraway and Schmidt (4)].

Hall and Nauda (1) provided a taxonomy that characterizes various approaches to research and development project selection. A common situation is that such methods generate a preferred subset of projects without considering implementation timing.

The problem of sequencing capacity expansion projects (SCEPS) is one of the most widely studied in the project sequencing literature (5,6). It is quantitatively based, requires specific information on demand, and yields decisions on the preferred projects and the corresponding sequence and timing. However SCEP and most other sequencing models cannot efficiently handle highly interdependent network effects.

The network design problem (NDP) approach has been applied to many network-related problems [e.g., see the papers by LeBlanc (7), Magnanti and Wong (8), and Janson et al. (9)]. The NDP model can consider the systemwide interactions among design decisions and analyze how design decisions affect the operations of a transportation network. However most existing NDP models are useful only for one-period decision making (i.e., project selection). The time dimension must be added to make NDP models suitable for project scheduling.

Akileswaran et al. (10) and Johnson et al. (11) have shown that SCEP and NDP models are fairly complex. Hence many researchers have used various heuristic solution methods [e.g., Poorzahedy and Turnquist (12)]. The most common difficulty encountered in any model is evaluation of network performance with respect to various changes. For a transportation network the traffic assignment model is frequently used to estimate the resulting total travel time. The computation time is quite large even for a network of moderate size.

The artificial neural network (ANN) has been studied as an alternative method for evaluating static network effects (13). When a time dimension is incorporated the ANN is shown to be an efficient prediction model for generating the multiperiod total travel times for any network changes (14). However it seems that applications of ANNs to transportation problems are just beginning (15). To date ANN research or practical applications in transportation engineering are still rare, although they are increasingly popular.

MPNDP MODEL

Given a transportation network and a set of improvement projects we try to find the optimal combination and schedule of projects that minimize the total discounted cost subject to relevant constraints. This is an MPNDP. We consider a planning horizon consisting of several equal discrete time periods (e.g., 1 year) and currently focus on capacity expansion of links (i.e., adding or improving a link).

The MPNDP model has the following features:

1. Both user and incremental supplier costs are considered in the objective function,
2. Periodic budgets are the only resource for construction, and the unspent portion may be rolled over into succeeding periods,
3. Project continuity is preserved, and
4. The resulting capacity changes in each period are specifically considered.

It is assumed that uncertainties about traffic demands and project costs may be disregarded (8). Only incremental supplier costs with respect to the null (i.e., existing) network will be counted. In this paper the projects are treated as immediately available to motorists in the first periods they are implemented. Another model considering project construction and possible traffic disruption has been proposed elsewhere (16). The solution framework discussed in the next section may be employed with a slightly different interpretation. However the computational burden would increase.

The following notation is used to present a mathematical formulation of the MPNDP model:

- A = set of links;
- B_h = budget for projects in period h ;
- C_{ah} = project cost of link a in period h , $a \in A$;
- $CRF_{r\tau}$ = capital recovery factor for discount rate r and τ periods;
- H = planning horizon;
- K_a = initial capacity of link a , $a \in A$;
- K_{ah} = capacity of link a in period h , $a \in A$;
- MAN_{ah} = maintenance cost of link a in period h , $a \in A$;
- N = set of nodes;
- $P(P')$ = set of links with (without) projects, $P' = A - P$;
- PVF_{rh} = present value factor for discount rate r in period h ;
- S_h = unspent budget in period h ; $S_0 = 0$;
- X = flow patterns on each link, $= (X_{ah})$ for $a \in A$ and $h \in H$;
- r = discount rate;
- x_{ah} = flow assigned on link a in period h ; $a \in A$ and $h \in H$;
- α_a, β_a = parameters of the travel time function on link a , $a \in A$;
- Δk_a = proposed additional capacity on link a , $a \in A$;
- ψ = unit cost of user travel time;
- μ_a = travel time at zero flow on link a , $a \in A$; and
- π_a = capital cost of link a , $a \in A$.

Assume m projects and τ periods are considered. Let V equal v_{ah} be the $m \times \tau$ decision matrix for $a = 1, \dots, m$, and $h = 1, \dots, \tau$. Each element in V is defined as follows:

$$v_{ah} = \begin{cases} 1 & \text{if project on link } \mathbf{a} \text{ is in service in period } \mathbf{h} \\ 0 & \text{otherwise} \end{cases}$$

Each matrix V represents a particular investment plan that specifies the preferred projects and implementation times. To preserve project continuity we require

$$v_{ah} \leq v_{a,h+1} \quad \forall \mathbf{a} \in P, h \in H \quad (1)$$

For example if link \mathbf{a} is to be improved in period 3, then the corresponding solution for link \mathbf{a} would be $v_{a1} = v_{a2} = 0$ and $v_{a3} = v_{a4} = \dots = v_{a\tau} = 1$. By thus defining decision variables capacity changes in each time period can be properly incorporated. Hence the corresponding average travel times can be accurately computed for the improved links. This new idea is presented here to reflect the effects of each improvement project.

The average travel time on link \mathbf{a} in period \mathbf{h} depends on project implementation and equilibrium flow. It is computed by

$$t_{ah}(x_{ah}, v_{ah}) = \mu_a [1 + \alpha_a (x_{ah} / K_{ah})^{\beta_a}] \quad (2)$$

where

$$K_{ah} = K_a + v_{ah} \Delta K_a \quad \forall \mathbf{a} \in P, h \in H \quad (3)$$

$$K_{ah} = K_a \quad \forall \mathbf{a} \in P', h \in H \quad (4)$$

By setting suitable initial link capacities both the link-adding and link-improving options can be handled simply by Equation 3. The initial capacity K_a may be assumed to be arbitrarily small for any possible new link, so that one unit flow will result in an extremely long travel time on this link. Therefore no traffic will be assigned to a yet nonexistent link. For existing links K_a is equal to its current physical capacity. Once this link is added or improved the second term on the right-hand side will ensure the addition of new capacity to the network. Then appropriate traffic volumes may be assigned accordingly.

The periodic project cost on link \mathbf{a} is computed by converting the capital cost to a periodic expenditure plus a maintenance cost in each period. Hence,

$$C_{ah} = \pi_a CRF_{r\tau} + MAN_{ah} PVF_{rh} \quad \forall \mathbf{a} \in P, h \in H \quad (5)$$

where

$$CRF_{r\tau} = \frac{r(1+r)^\tau}{(1+r)^\tau - 1} \quad (6)$$

$$PVF_{rh} = 1 / (1+r)^h \quad (7)$$

In principle the periodic maintenance cost may depend on the age or utilization rate of the facility as discussed by Markow (17) and Fwa et al. (18). However practically reliable results are still underdeveloped (19). Hence MAN_{ah} is assumed to be a fixed fraction of the project capital cost in the present study.

The system cost is defined for each period as the sum of user travel time costs and project costs:

$$\text{System cost in period } h = \sum_{a \in A} [\Psi x_{ah} t_{ah}(x_{ah}, v_{ah}) + v_{ah} C_{ah}] \quad (8)$$

It is clear that the system cost depends not only on the project implementation decisions but also on the traffic flows on each link. Furthermore the flow patterns will be updated according to the projects selected up to the current period. There seems to exist a hierarchy for this problem. A higher-level position for the decisions on projects seems appropriate. Given the decision variables v_{ah} the equilibrium flow assignment may be processed at the lower level. Consequently MPNDP is expressed by two subproblems at different levels.

We now define the solution set fl for MPNDP as

$$\Omega = \{V_i = (v_{ah}) | v_{ah} = 0 \text{ or } 1, i = 1, 2, \dots, (\tau + 1)^m\} \quad (9)$$

A total number of $(\tau + 1)^m$ possible solutions is included in Ω for the corresponding MPNDP.

The MPNDP consists of two parts, namely the network priority program problem (NPPP) in the upper level and the periodic network equilibrium problem (PNEP) in the lower level. The NPPP is formulated below as a nonlinear mixed-integer program subject to constraints representing periodic funds available and project continuity. The NPPP formulation is:

$$\text{Minimize } Z = \sum_{v \in \Omega} \sum_{a \in A} [\Psi T_{ah}(x_{ah}^*, v_{ah}) + v_{ah} C_{ah}] \quad (10)$$

$$\sum_{a \in P} (v_{ah} - v_{a,h-1}) \pi_a / PVF_{r,h-1} + S_h - S_{h-1} / PVF_{r,1} = B_h \quad \forall h \in H \quad (11)$$

$$S_h \geq 0 \quad \forall h \in H \quad (12)$$

$$v_{a,h-1} \leq v_{ah} \quad \forall a \in P, h \in H \quad (13)$$

$$v_{ah} = 0 \text{ or } 1 \quad \forall a \in P, h \in H \quad (14)$$

In Equation 10 x_{ah}^* is the optimal solution of the following network equilibrium problem in period h for any given feasible decision matrix V .

The PNEP formulation is:

$$\text{Minimize}_x Z' = \sum_{a \in A} \int_0^{x_{ah}} t_{ah}(u, v_{ah}) du \quad (15)$$

subject to

$$\sum_{k \in K_{rsh}} f_{kh}^{rs} = q_{rsh} \quad \forall r, s \in R_h \quad (16)$$

$$x_{ah} = \sum_{r, s \in R_h} \sum_{k \in K_{rsh}} f_{kh}^{rs} \delta_{akh}^{rs} \quad \forall a \in A \quad (17)$$

$$f_{kh}^{rs} \geq 0 \quad \forall k \in K_{rsh}, r, s \in R_h \quad (18)$$

where

- K_{rsh} = set of paths connecting origin-destination (O-D) pair r - s in period h , for $r, s \in R_h$;
- R_h = set of origins and destinations in period h , $R_h \subseteq N$;
- f_{kh}^{rs} = flow on path k connecting O-D pair r - s in period h ;
- q_{rsh} = trip rate between O-D pair r - s in period h ; and
- δ_{akh}^{rs} = 1 if link a is on path k between O-D pair r - s in period h and 0 otherwise.

The bilevel structure of the MPNDP model is similar to those presented by LeBlanc and Boyce (20) and Bard (21). However the proposed model is more realistic since the improvements are considered for the different demands and corresponding user behaviors in each period throughout

the planning horizon. On the other hand this model is considerably more difficult to solve because of the extensive and complex interactions between users and planners.

SOLUTION METHOD

Considering the project continuity constraint there are only $\tau + 1$ possible decisions for each row (i.e., for each individual project) in the decision matrix. These cases may be represented by summing up the values of decision variables in the same row. Hence only a row sum variable v_a is needed for any possible implementation of project a . Consequently the row sum vector may be appropriately constructed with the following definition to replace the decision matrix V :

$$RS = (v_1, v_2, \dots, v_m)^T \quad (19)$$

$$v_a = \tau + 1 - \sum_{h=1}^{\tau} v_{ah} \quad \forall a \in P \quad (20)$$

In Equation 19 T stands for the transpose of a vector.

The row sum variable is a convenient representation since each value corresponds to a decision on project selection and scheduling. Then we may modify Equation 9 as

$$\Omega = \{RS_i | i = 1, 2, \dots, (\tau + 1)^m\} \quad (21)$$

Note that with Equations 19 through 21 all elements in the set Ω already implicitly fulfill the project continuity constraint. Hence only the budget constraint remains to be satisfied in the solution procedure.

It has been shown that an efficient project sequence is quite helpful in the solution process [e.g., Erlenkotter (22), Janson and Husaini (23), and Martinelli (24)]. It is usually obtained by ranking the relative effects of projects on the system. A good initial project sequence can speed up the proposed solution method. The initialization criterion used here is the saving/cost ratio of each individual project.

To solve the MPNDP a branch-and-bound (BB) procedure along with an ANN model is developed. The proposed procedure can cost-effectively evaluate the resulting system cost for each solution considered and screen inferior solutions to quickly obtain the optimal solution.

ANN Model

The motivation and justification of using the ANN approach is its small predictive error as well as its reasonable computational burden. In particular when only the total travel time in a

transportation network is needed a relatively simple ANN model may serve as a proxy for the conventional traffic assignment model (14). However several specific choices must be made for the training parameters.

The ANN model is constructed to compute the system equilibrium (SE) user travel times, taking into consideration the effects of project selection, scheduling, and different demands over time. The desirable feature of the ANN approach is that, after the ANN is trained, it may be repeatedly used for any analysis on the MPNDP, in which each replication requires very little computation time. The ANN approach is especially suitable for relatively large transportation networks in which long computation times are usually required for traffic assignments. Some relevant discussion and validation are provided by Wei (16).

BB Procedure

Considering various factors in the transportation network improvement problem, a preliminary conclusion is that lower total system costs tend to be associated with the earlier implementation of projects. Hence the objective function of NPPP is roughly a U-shaped curve skewing to smaller values of row sum variables. This property is particularly important on capacitated networks where congestion effects increase user travel time exponentially. The proposed BB method is mainly based on this observation, and detailed discussions may be found in Wei (16).

With the initial project sequence a synthesized branch rule is developed and the ANN model is activated whenever a lower bound (LB) is needed in the solution process. The conventional traffic assignment is used to estimate the user equilibrium (UE) user travel times for each complete solution. On the basis of the branch rule the proposed BB method would generate a tree with as many levels as the number of row sum variables (i.e., number of projects). Hence the level index L is also used as the project index.

To monitor the progress of the BB method a list containing the branch indexes in descending order is needed. Information about the new branch is added to the top of the branch index list. Each branch index is associated with a partial solution or a complete solution when the level index is equal to m . In any case the branch with the largest index is at the top of the list and will be processed first. As a general rule the indexes of branches from the same predecessor should be labeled in the reverse order of the assigned values for the variables under current consideration. The branch index and associated information will be removed from the list after further partitioning or fathoming is accomplished.

The core of the proposed BB method is to choose the best possible solution (BPS) for each branch, given the decision on already specified projects. Since each branch represents a number of possible solutions, the intelligently derived BPS would sufficiently reflect the goodness of the associated solutions. Such a task is accomplished by estimating and updating the earliest implementation times (EITS) of all unspecified projects.

The EIT of project a , h_a , is the smallest time index in which project a may be implemented without violating the relevant constraints as well as the schedule of already specified projects. For each partial solution updating of EITs is equivalent to choosing the smallest values for free row sum variables according to the fixed values of other variables. The proposed procedure is described below.

At level 0 (i.e., root of the BB tree) the EITs of all projects are verified by

For partial solutions at level $L > 0$ the first L projects have been specified to have fixed values. The remaining $m-L$ variables are free, and their updated EITs corresponding to those fixed

$$h'_a = \min \left\{ j \left| \sum_{h=1}^j B_h PVF_{r,h-1} \geq \pi_a \right. \right\} \quad \forall a \in P \quad (22)$$

variables must be decided. The largest value among the already specified variables is identified by

$$v'_L = \begin{cases} \max \{v_i | i \leq L\} & \text{if some } v_i \leq \tau \\ 0 & \text{otherwise} \end{cases} \quad (23)$$

v'_L indicates the last period for accumulating available budget. When v'_L is zero the projects specified so far are not to be implemented and the budget is not used at all. Thus the EITs of the free variables are set equal to h'_a , obtained in Equation 22. For nonzero v'_L the remaining budget is then obtained by subtracting the construction costs of the already implemented projects.

The appropriate EITs for free variables are determined by one of the following conditions:

1. If the remaining budget is larger than any construction cost of the free projects, the prevailing unspecified projects may be also implemented before period v'_L without exceeding the budget limit. Thus when Equation 24 holds for any free project the corresponding EIT is set equal to the EIT obtained in its predecessor node.

$$\sum_{h=1}^{v'_L} B_h PVF_{r,h-1} - \sum_{a=1}^L \pi_a \geq \pi_i \quad L < i \leq m \quad (24)$$

2. Otherwise the EIT of free project i is obtained by

$$h_i = \min \left\{ j \left| \sum_{h=1}^j B_h PVF_{r,h-1} - \sum_{a=1}^L \pi_a \geq \pi_i \right. \right\} \quad L < i \leq m \quad (25)$$

Note that the EIT of each unspecified project obtained is thus based on the budget relaxation proposed by Wei (16). This is to ensure the feasibility of already chosen projects and the achievement of lower costs from all unselected projects. Thus the greatest contribution that each

project may yield to the system is obtained on the basis of the currently established network. Such budget relaxation is also desirable to reduce the problem complexity since the not yet considered projects will compete for the remaining budget. As a result the complete solutions in the BB tree are always budget feasible.

With the above treatments new branches can be created rapidly and more partial solutions can be examined for their system effectiveness. Since the ANN model is fairly efficient the lower bound is quite tight and the overall solution process is very fast.

Algorithmic Procedures

The complete solution algorithm for the MPNDP is condensed as follows:

- Step 0: *Preprocess.* Presort projects according to their relative system effectiveness and assign the project index in that order.
- Step 1: *ANN Training.* Train the ANN by using the methods discussed by Wei (16).
- Step 2: *Initialization.*
 - a.. Set L equal to 0.
 - b. Compute initial upper bound (UB) equal to Z_{UE} under the current network configuration.
 - c. Compute the EITs at level L for all projects by using Equation 22.
- Step 3: *Branching.*
 - a. Set L equal to $L + 1$.
 - b. First, for $L < m$ partition v_L according to the updated EIT, assign branch indexes, generate partial solutions, and put this information on the branch index list. Second, for L equal to m partition v_m according to the updated EIT, assign branch indexes, generate complete solutions, and put this information on the branch index list.
- Step 4: *Bound Computation.*
 - a. Pick the first branch and the associated partial solution from the list revised in Step 3.
 - b. If L is equal to m go to Task A. Otherwise obtain the updated EITs for free variables by using Equation 24 or 25, estimate the SE total travel time for the corresponding BPS with the trained ANN, and compute the LB.
- Step 5: *Comparison.*
 - a. For LB greater than or equal to UB fathom this solution and go to Task B.
 - b. For LB less than UB go to Step 3 if L is less than m ; otherwise store this incumbent solution, set UB equal to LB, and go to Task B.
- Task A: *Computing Z_{UE} for Complete Solutions.* For the complete solution perform UE traffic assignment and compute Z_{UE} under the current project schedule. Set LB equal to Z_{UE} and go to Step 5.
- Task B: *Checking the Branch Index List*
 - a. For L equal to 1:
 - If there is no branch at the same level, stop the BB process; the latest incumbent solution is the optimal solution.
 - Otherwise go to Step 4.
 - b. For L greater than 1:
 - If there is no branch at level L go to level $L - 1$.

- If there is no branch at level $L - 1$, set L equal to $L - 1$ and go to Task B; otherwise set L equal to $L - 1$ and go to Step 4.
- Otherwise go to Step 4.

For a three-project, 5-year case discussed by Wei (16) the BB solution process is shown in Figure 1. Note that because of the relatively small problem size the total costs of possible solutions are quite close. Hence quite a few complete solutions are evaluated at the lowest level. As shown in the next section the proposed solution method is very efficient and only a few complete solutions need to be evaluated when a practical problem size is considered.

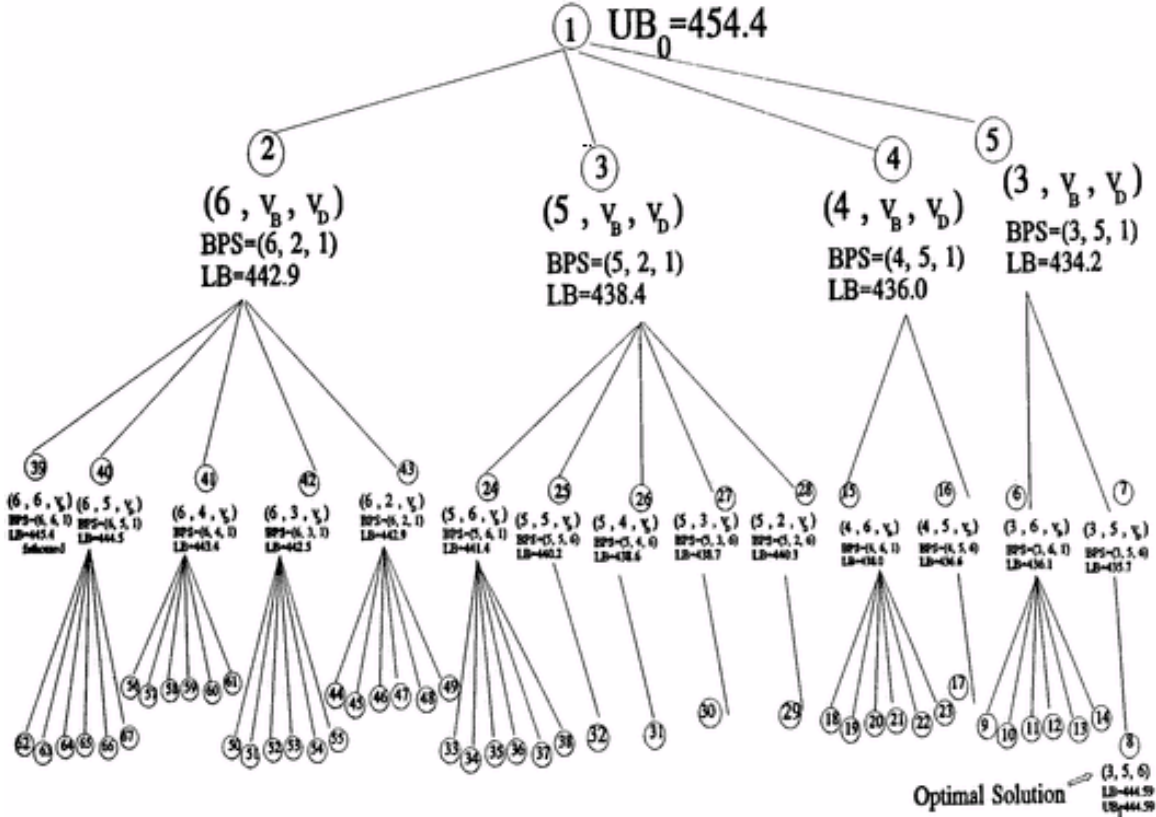


FIGURE 1 Illustration of proposed BB method.

NUMERICAL EXAMPLE

A realistic problem is used in this section for demonstrating the solution method proposed for the MPNDP model. The relevant information was provided by the Maryland State Highway Administration for a related study (25). The characteristics of this illustrative problem are practical enough that it can be used to validate the usefulness of the proposed methodology for real-world problems.

To alleviate future congestion in Calvert County, Md., five projects are considered, as shown in Figure 2. Projects X , Y , and Z add one more traffic lane to the associated links in each direction. Alternatively projects P and Q provide bypass routes for most of the congested areas. The bypass routes are assumed to be two-lane, two-way highways.

These five projects, if all are completed, would greatly relieve future traffic congestion. However this is hardly possible because of limited funds. In addition since two parallel routes exist for most of the congested areas the project effects tend to overlap. Hence installing two parallel projects simultaneously is unlikely to be efficient.

Projects are considered for a 12-year planning horizon, from the years 1999 to 2010. All cost and saving computations will be based on the resent value in the year 1999. The primary factors used for these computations are listed in Table 1. According to Equation 21 there are 13^5 (371,293) possible combinations of projects and schedules to be evaluated, which is not a trivial task.

TABLE 1 Parameters Used for Methodology Demonstration

Item	Value
Unit Value of Time (\$/Veh-hour)	10.0
Peak Hour to Day Ratio	0.15
Day to Year Ratio	1/300
Planning Horizon (Years)	12
Interest Rate (%)	6
New Construction Cost (Million \$/Lane mile)	4.5
Roadway Widening Cost (Million \$/Line mile)	3.0
Ratio of Overhead and Other Costs to Construction Cost	20
Annual Maintenance to Construction Cost Ratio	1.55

To set up an MPNDP for this case a number of preliminary analyses are conducted. Without improvements it is found that the overall average speed is reduced from 38.25 mph in the year 1999 to 20.86 mph in the year 2010. Hence significant improvements on this highway system are desirable to preserve a reasonable level of service. The effects and capital costs of each individual project are listed in Table 2. The last column in Table 2 shows the cost-effectiveness rank of each individual project and constitutes the solution of many scheduling methods.

Table 2 also provides some information about project combinations. In particular the total user travel time is almost halved and the average travel speed is almost preserved at the year 1999 level if all projects are implemented. Nevertheless the corresponding travel time savings are notably less than the sum of individual ones. This explicitly indicates the interdependencies among various projects.

According to the factors listed in Table 1 the total cost of the null system is \$2,371 million. Assuming that the available budget is \$15 million/year, this test problem is solved with a trained ANN and the proposed BB method. The solution process takes only a few seconds of central processing unit time on a 486-based personal computer. The optimal scheduling solution for (P, X, Y, Z, Q) is (3, 11, 5, 5, 13), with a total cost \$1,743 million. The costly project Q is not considered for implementation, although its time savings is among the highest. The system cost savings that would result from this implementation plan are \$628 million (or 26 percent of the null alternative) for the 12-year planning horizon.

It is helpful to justify the usefulness of the proposed methodology by comparing its results with those of the scheduling decision obtained on the basis of independent project effects. The approach is to use the independent sequence shown in Table 2 and to determine the project implementation times that lead to the minimum total cost. Given the same conditions discussed above the optimal scheduling solution is (3, 4, 7, 11, 13) and has a total cost \$1,841 million. It is clear that a better solution, with \$98 million of additional savings, is found by considering project interdependencies.

TABLE 2 Project Effects and Ranking in Year 2010

Project	Length miles	Project Cost ¹ (A)	Total UTT ²	Speed ³ MPH	UTT Saving (B)	B/A Ratio
Null	-	-	15500	20.86	-	-
p	5.40	29.2	12242	26.32	3258	112
x	5.98	21.5	13621	23.69	1879	87
y	10.27	37.0	12359	26.07	3141	85
z	8.38	30.2	13937	23.26	1563	52
Q	12.88	69.6	12351	26.02	3149	45
Combination						
XYZ	24.63	88.7	9000	35.73	6500	73
PQ	18.28	98.8	10956	28.99	4544	46
ALL	42.91	187.5	8521	36.82	6979	37

¹Million dollars

²User equilibrium travel time, veh-hours/peak-hour

³Average peak hour speed at network level, miles/hour

The effects of various budget levels are analyzed. The approach is to restrict the annual budget so that the present value of total budgets is a certain fraction of the total project costs. Six budget levels, from 50 to 100 percent of total project costs, are considered, and the results are shown in Table 3. It is interesting to note that optimal solutions for different budget levels yield similar improvement effects over the null system, as shown in the last column of Table 3. However the optimal scheduling solutions and the corresponding total system costs are quite different.

It is found that for lower budget levels (e.g., 50 and 60 percent) improvements on existing links are preferred since the associated costs are usually lower. The new bypass routes are either deferred or not considered for installation. If, however, the budget is insufficient (e.g., 70 percent or higher) new links may be added in the early stages.

Table 3 also provides information on the processes of the proposed BB method, that is, the numbers of nodes created and the numbers of feasible solutions evaluated. Since the nodes represent both partial and complete solutions generated throughout the solution procedure, this information indicates that the proposed BB method is fairly effective. The infeasible or inferior solutions are screened out efficiently because of the specially designed branching and bounding rules. Only a small fraction of possible solutions must be evaluated. This demonstrates the highly desirable property addressed in the previous section. Additionally the information about BB nodes seems to

indicate that the proposed method is best suited for budget levels of between 80 and 100 percent of total project costs.

TABLE 3 Results for Various Budget Levels

Budget Level ¹	Annual Budget	BB Nodes ²	# Feasible Solutions	Optimal Solution	Total Cost	Improvement Over Null ³
50%	10.5	463	9	(11,3,4,13,13)	1722	27%
60%	12.5	565	7	(10,2,4,13,13)	1687	29%
70%	15.0	120	4	(3,11,5,5,13)	1743	26%
80%	17.0	75	3	(8,2,3,9,13)	1726	27%
90%	19.0	49	2	(2,3,8,5,13)	1729	27%
100%	21.1	42	2	(2,3,5,7,13)	1747	26%

¹Total project cost = 187.5

²Nodes created in the branch-and-bound solution process

³Total cost of null alternative = 2371

POTENTIAL APPLICATIONS

The MPNDP model and solution method proposed in this paper is especially designed for prioritizing interrelated projects in transportation networks. Below several potential applications of the proposed MPNDP model are discussed.

Application in Highway Maintenance Planning

Conventional highway maintenance planning tends to neglect the impacts on roadway users (19). Hence the resulting maintenance plan is rarely the best conceivable. The combined costs of highway maintenance and traffic operations must be considered for proper maintenance planning. In particular when major rehabilitation is undertaken the influence on existing traffic patterns is fairly significant.

Various maintenance alternatives may be treated as possible projects that recover the network performance to different levels. Then the traffic assignment model may be used to estimate the aggregate utilization of the roadway system. Consequently the mutual influences between the user and the facility can be properly taken into account. For example, the actual deterioration would depend on route selection by drivers, which in turn affects maintenance needs.

HOV Lanes and IVHS Applications

The proposed MPNDP can be applied to evaluate various traffic improvement plans. For example it can be used for determining the suitable stages for introducing high-occupancy-vehicle (HOV) lanes in different locations. With small additional efforts the proposed methodology may also be used to plan advanced transportation systems, for example, intelligent vehicle-highway systems (IVHSs).

A critical issue in these applications is assessment of the traffic pattern changes owing to HOV lanes or various IVHS technologies. In particular only a fraction of conventional users and facilities will be affected. Special traffic assignment models are thus needed to deal with vehicles with various occupancies or equipment.

With such information proper samples for ANN training can be generated according to the plans under consideration. Then the MPNDP for implementing HOV lanes or IVHS technologies within a certain horizon can be formulated and solved by the proposed BB method.

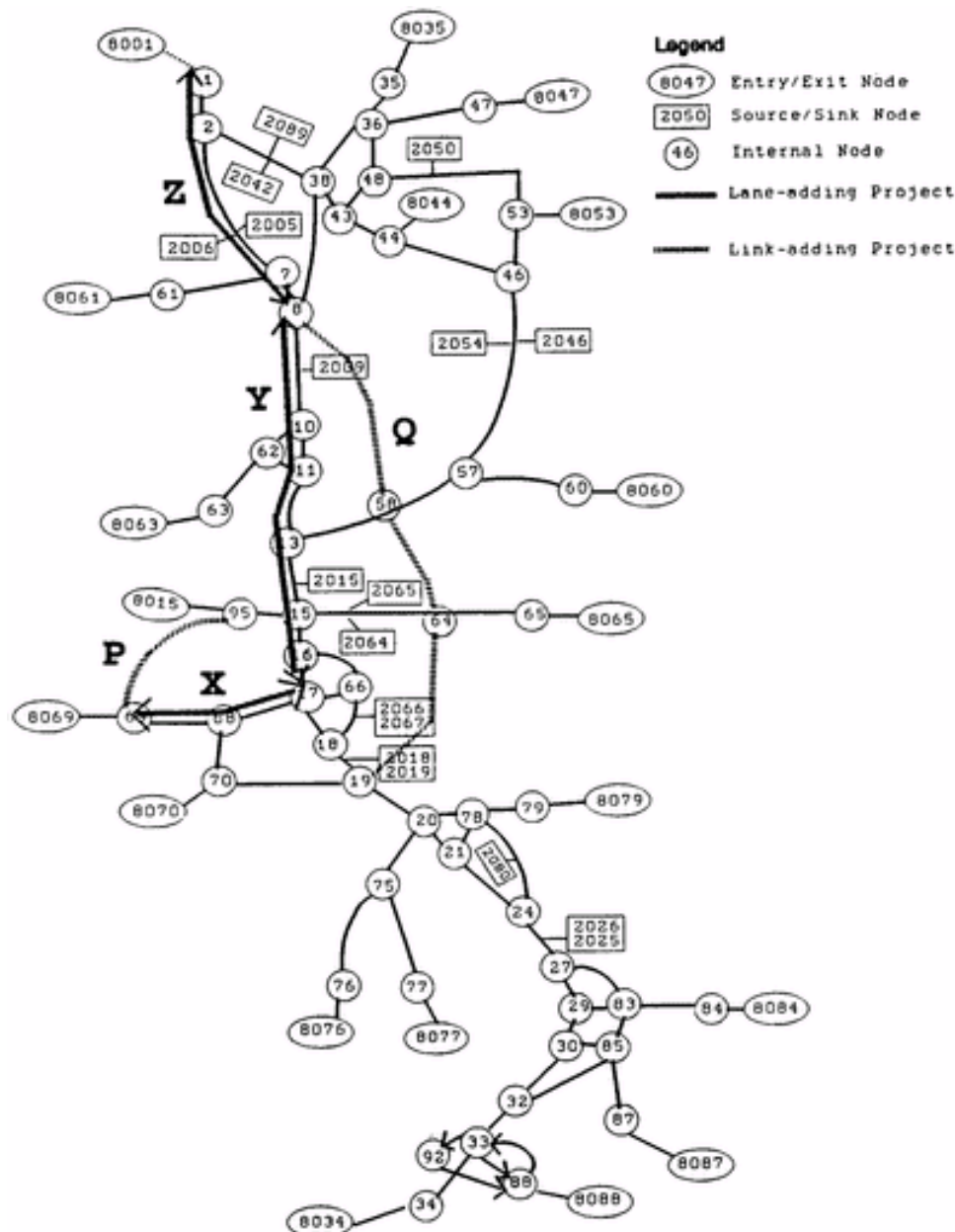


FIGURE 2 Proposed projects for Calvert County, Md., by year 2010.

CONCLUSIONS

Selecting the optimal project combination and implementation timing is very important for transportation systems. This problem tends to be fairly difficult since complex project interrelations often exist. The main drawbacks of most existing methods are long computation times and neglect of conditions in the intermediate period. The latter may lead to inappropriate solutions. A model for multiperiod transportation network priority programming was developed in the study described here. This model has the desirable features of simultaneously determining the best combination of projects and schedules. The proposed model is more realistic than others since the improvements are considered for the different demands and corresponding user behaviors in each time period throughout the planning horizon.

To solve the MPNDP a BB procedure is specifically designed. The ANN approach is adapted to compute the resulting user travel times, taking into consideration the effects of project selection, scheduling, and different demands over time. The overall solution method can evaluate possible solutions very cost-effectively and can screen out many inferior solutions to save computational efforts. The numerical examples show that only a small fraction of possible solutions must be evaluated and the proposed BB method seems to be especially fast for budget levels of between 80 and 100 percent of total project costs.

The MPNDP model may be considered for many other network-related problems in which interrelated projects must be scheduled.

Several conceivable extensions of the proposed methodology are worth pursuing, for example, highway maintenance planning, suitable stages for HOV lanes in different locations, and the transition timing of various IVHS technologies.

REFERENCES

1. Hall, D. L., and A. Nauda. An Interactive Approach for Selecting IR&D Projects. *IEEE Transactions on Engineering Management*, Vol. 37, No. 2, 1990, pp. 126-133.
2. Fox, G. E., et al. Economic Models for R&D Project Selection in the Presence of Project Interactions. *Management Science*, Vol. 30, 1984, pp. 890-902.
3. Gomes, L. F. A. M. Modeling Interdependencies Among Urban Transportation System Alternatives Within a Multicriteria Ranking Framework. *Journal Advanced Transportation*, Vol. 24, No. 1, 1990, pp. 77-85.
4. Carraway, R. L., and R. L. Schmidt. An Improved Discrete Dynamic Programming Algorithm for Allocating Resources among Interdependent Projects. *Management Science*, Vol. 37, No. 9, 1991, pp. 1195-1200.
5. Neebe, A. W. and M. R. Rao. The Discrete-Time Sequencing Expansion Problem. *Operations Research*. Vol. 31, No. 3, 1983, pp. 546-558.
6. Neebe, A. W., and M. R. Rao. Sequencing Capacity Expansion Projects in Continuous, Time. *Management Science*, Vol. 32, No. 11, 1996, pp. 1467-1479.

7. LeBlanc. L. J. An Algorithm for the Discrete Network Design Problem. *Transportation Science*, Vol. 9, 1975, pp. 183-199.
8. Magnanti. T. L., and R. T. Wong. Network Design and Transportation Planning: Models and Algorithms. *Transportation Science*, Vol. 18, 1984. pp. 1-55.
9. Janson. B. N., L. S. Buckets, and B. E. Peterson. Network Design Programming of U.S. Highway Improvements. *ASCE Journal of Transportation Engineering*. Vol. 117, No. 4, 1991.
10. Akileswaran. V. G. B. Hazen, and T. L. Morin. Complexity of the Project Sequencing Problem. *Operations Research*, Vol. 31, No. 4, 1983, pp. 772-778.
11. Johnson. D. S., J. K. Lenstra, and A. H. G. Rinnooy Kan. The Complexity of the Network Design Problem. *Networks*, Vol. 8, No. 4, 1978, pp. 279-285.
12. Poorzahedy, H., and M. A. Turnquist. Approximate Algorithms for the Discrete Network Design Problem. *Transportation Research*, Vol. 16B, 1982, pp. 45-55.
13. Xiong, Y, and J. B. Schneider. Transportation Network Design Using a Cumulative Genetic Algorithm and a Neural Network. Presented at 71st Annual Meeting of the Transportation Research Board, Washington, D.C., 1992.
14. Wei, C. H., and P. M. Schonfeld. An Artificial Neural Network Approach for Evaluating Transportation Network Improvements. *Journal of Advanced Transportation*, Vol. 27, No. 2, 1993, pp. 129-150.
15. Faghri, A., and J. Hua. Evaluation of Artificial Neural Networks Applications in Transportation Engineering. Presented at 71st Annual Meeting of the Transportation Research Board, Washington, D.C., 1992.
16. Wei, C. H. *Priority Programming for Transportation Networks using Artificial Neural Networks*. Ph.D. dissertation. University of Maryland, 1993.
17. Markow, M. J. *Demand Responsive Approach to Highway Maintenance and Rehabilitation*. Executive Summary, Report DOT/OST/ P34/87/052, June 1985.
18. Fwa, T. F., et al. Highway Routine Maintenance Programming at Network Level. *ASCE Journal of Transportation Engineering*, Vol. 114, No. 5, 1988, pp. 539-554.
19. Wei, C. H., and P. M. Schonfeld. *The Combined Cost of Highway Maintenance and Traffic Operations*. Report MD-93/02. Maryland State Highway Administration, 1992.
20. LeBlanc, L. J., and D. E. Boyce. A Bilevel Programming Algorithm for Exact Solution of the Network Design Problem with User-Optimal Flows. *Transportation Research*, Vol. 20B, 1986, pp. 259-265.
21. Bard, J. F. An Algorithm for Solving the General Bilevel Programming Problem. *Mathematics of Operations Research*, Vol. 8, 1983, pp. 260-272.
22. Erlenkotter, D. Sequencing of Interdependent Hydroelectric Projects. *Water Resources Research*, Vol. 9, No. 1, 1973, pp. 21-27.
23. Janson, B. N., and A. Husaini. Heuristic Ranking and Selection Procedures for Network Design Problems. *Journal of Advanced Transportation*, Vol. 21, No. 1, 1987, pp. 17-46.

24. Martinelli, D. *Investment Planning of Interrelated Waterway Improvement Projects*. Ph.D. dissertation. University of Maryland, College Park, 1991.
25. Wei, C. H., and P. M. Schonfeld. *Priority, Programming for Congested Transportation Networks*. Report MD-92/03. Maryland State Highway Administration, 1992.

Publication of this paper sponsored by Committee on Transportation Programming, Planning, and Systems Evaluation.